Theoretical and Computational Studies of Protein Folding Energy Landscapes

Dissertation
zur Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

Mathematisch-naturwissenschaftlichen Fakultät der
Universität Zürich

von
Francesco Rao
aus
Italien

Promotionskomitee

Prof. Dr. Amedeo Caflisch
Prof. Dr. Ben Schuler

Zürich 2005

List of publications

The protein folding network
F. Rao and A. Caflisch
J. Mol. Biol. 342, 299-306 (2004)

Local modularity measure for network clusterizations
S. Muff, F. Rao, and A. Caflisch
Phys. Rev. E 72, 056107 (2005)

Estimation of protein folding probability from equilibrium simulations
F. Rao, G. Settanni, E. Guarnera and A. Caflisch
J. Chem. Phys. 122, 184901 (2005)

$\Phi-$value analysis by molecular dynamics simulations of reversible folding
G. Settanni, F. Rao and A. Caflisch
Proc. Natl. Acad. Sci. 102, 628-633 (2005)

Replica exchange molecular dynamics simulations of reversible folding
F. Rao and A. Caflisch
J. Chem. Phys. 119, 4035 (2003)

Replica exchange molecular dynamics simulations of amyloid peptide aggregation
M. Cecchini, F. Rao, M. Seeber and A. Caflisch
J. Chem. Phys. 121, 10748-10756 (2004)

Summary

The present thesis is concerned with the development and application of three novel approaches for the generation and the analysis of energy landscapes characterizing protein folding. In the past decade, the energy landscape paradigm has emerged as a consistent and useful framework for the study of many aspects of protein folding leading to the so-called funnel picture for folding. Projections of the energy landscape to order parameters should give a simple and synthetic view of the states and energetic barriers of the system. Free-energy projections are used to estimate the stability and in some cases also the kinetics of the states involved in the folding process. However, this simplistic picture revealed some hidden assumptions on order parameter(s) that in many cases are not fulfilled.

First, a novel approach for the study of the states of a protein, based on *complex networks*, is introduced. Conformations visited during a molecular dynamics (MD) simulation and the transitions between them are the nodes and the links of the network, respectively. The network represents the multidimensional free-energy landscape and does not require any projections to arbitrarily chosen order parameters. The network approach revealed a complex scenario of minima, basins and super-basins of attraction in the free-energy landscape. As a first application, a structure peptide, namely beta3s, was investigated by MD simulations and network analysis. Both low-enthalpy/low-entropy and high-enthalpy/high-entropy basins were found in the denatured state. Interestingly, the native state of beta3s is stabilized not only enthalpically but also entropically, the latter with respect to a kinetic trap. In the network framework, free-energy basins are defined as subgraphs (called communities or clusters) of the network where nodes are highly connected. Although many communities detection algorithms exist in literature, it is not yet clear how to judge the quality of a clusterization. A new measure to assess this problem, called *goodness deviation*, is proposed in the thesis.

Second, kinetics of protein folding are strongly connected to the topography and the organization of the energy landscape. The elusive nature of the transition state ensemble (TSE) doesn't allow an easy and clear picture of the folding energy barrier. A common technique, widely used in literature, to char-

acterize the TSE is the computation of the folding probability p_{fold} which is the probability of a structure to fold before unfolding. Even if p_{fold} is a useful quantity for transition state analysis, it requires a very large computational effort which in practice limits its application to the analysis of small sets of data. The kinetic homogeneity of structurally similar snapshots suggests the use of a statistical approach for the analysis of the kinetic properties of a peptide conformation. In this thesis, a novel approach for the computation of the folding probability, called $cluster - p_{fold}$, allows the approximate estimation of the p_{fold} for every structure sampled along an MD simulation. The application of the method on the beta3s peptide has shown the presence of a broad and heterogeneous TSE as well as two average folding pathways. Interestingly, these findings are in agreement with the network analysis performed on the peptide in our former studies. The negligible computational demand for the calculation of the $cluster - p_{fold}$ allowed the analysis of the TSE for a large set of beta3s mutants and the calculation of the $\Phi-$values. Given the huge amount of simulation time ($\sim 0.65ms$), this analysis would have been impossible with traditional methods.

Third, all constant temperature simulations presented in this thesis were run at temperatures higher than the physiological temperature. It is worth noting that the height of energy barriers is proportional to $e^{-\frac{\Delta E}{kT}}$, where k and T are the Boltzmann constant and temperature, respectively. For this reason, MD simulations ran at physiological temperatures get trapped very easily in local minima and are not able to correctly sample the energy landscape in a reasonable amount of computational time. In this thesis, the *replica-exchange method* (REM), is used to simulate the folding-unfolding process of the beta3s peptide at different temperature values ranging from $275K$ to $465K$. An extensive comparison to constant temperature simulations shows that REM is able to correctly sample the energy landscape at physiological temperatures which is not possible by conventional MD. One disadvantage of the method is that it cannot be used for the analysis of the kinetics. REM simulations were also used to study the early stages of peptide aggregation of an amyloidogenic peptide from the yeast prion protein sup35.

Zusammenfassung

Die folgende Dissertation befasst sich mit der Entwicklung und Anwendung dreier neur Ansätze für die Erzeugung und Untersuchung von Energy Landscapes mit dem Ziel, die Proteinfaltung zu charakterisieren und besser zu verstehen. Das Paradigma der Energy Landscapes hat sich in den letzten zehn Jahren bei der Untersuchung des Faltungsverhaltens von Proteinen bewährt. Dieses Paradigma führt dazu, dass die Energy Landscape eines Proteins vereinfacht als Trichter dargestellt werden kann. Projektionen dieser Energy Landscape auf Ordnungsparameter, wie zum Beispiel die Projektion auf die freie Energie zur Untersuchung von Stabilität, sollten eine vereinfachte Sicht der Zustände und Energiebarrieren wiedergeben, wobei oftmals versteckte Annahmen getroffen werden, die nicht erfüllt sind.

Basierend auf der Theorie *komplexer Netzwerke* wird als erstes ein neur Ansatz zur Untersuchung der Proteinzustände vorgestellt. Dabei sind die Konformationen, welche im Verlauf einer molecular dynamics (MD) Simulation angenommen werden, die Knoten und die zeitlichen Uebergänge die Verbindungen. Dieses Netzwerk repräsentiert die vollständige multi-dimensionale Energy Landscape ohne Projektion auf willkürlich gewählte Ordnungsparameter. Die Anwendung von Netzwerken enthüllt ein komplexes Zusammenspiel von Minima, Basins und Super-Basins, die in der Energy Landscape die Rolle von Attraktoren spielen. Als erstes wird die MD-Simulation eines strukturierten Peptides (Beta3s) untersucht. Dabei werden Basins mit kleiner Entropie/kleiner Enthalpie, sowie Basins mit grosser Entropie/grosser Enthalpie gefunden. Hierbei ist interessant, dass Beta3s im gefalteten Zustand (im Unterschied zu einer kinetischen Trap) nicht nur enthalpisch, sondern auch entropisch stabilisiert wird. Free Energy Basins entsprechen Subgraphen (Communities) des gesamten Netzwerks, wobei die Knoten in einen Basin dicht miteinander verbunden sind. Das Finden dieser Communities und die Validierung einer Unterteilung des gesamten Netzwerkes in solche ist trotz der Fülle von existierenden Algorithmen noch nicht zufriedenstllend gelöst. In dieser Dissertation wird dazu ein Mass, die *Goodness Deviation*, vorgeschlagen.

Zweitens ist die Kinetik der Proteinfaltung ebenfalls eng mit der Topogra-

phie von Energy Landscapes verknüpft. Die schwer fassbare Natur des Transition State Ensemble (TSE) erlaubt kein einfaches Bild für Faltunsübergänge. Zur Charakterisierung des TSE wird oft die Faltungswahrscheinlichkeit p_{fold} einer Proteinstruktur benutzt. Obwohl p_{fold} für die Analyse von Faltungsübergängen sehr nützlich ist, erfordert dessen Berechnung einen grossen rechnerischen Aufwand, was die praktische Anwendungen limitiert. Die kinetische Homogenität von strukturell ähnlichen Snapshots erlaubt jedoch einen statistischen Zugang für die Analyse einzelner Konformationen. Die Idee des sogenannten *Cluster-*p_{fold} ermöglicht es in dieser Dissertation, p_{fold} für Strukturen einer MD-simulation approximativ zu berechnen. Die Anwendung dieser Methode auf Beta3s zeigt ein breites, heterogenes TSE, sowie zwei Haupt-Pathways. Diese Ergebnisse decken sich interessanterweise mit der Netzwerk-Analyse aus vorhergehenden Untersuchungen. Der marginale Rechenaufwand für den *Cluster-*p_{fold} ermöglichte die Analyse des TSE von vielen Beta3s-Mutanten und die Berechnung von Φ-Werten. Mit herkömmlichen Methoden wäre dies für unsere riesige Menge Simulationsdaten ($\approx 0.65ms$) unmöglich gewesen.

Drittens sind in dieser Dissertation alle Simulationen mit konstanter Temperatur bei einer Temperatur T gemacht worden, welche höher ist als die physiologische Temperatur. Es sei erwähnt, dass die Höhe der Energie-Barrieren proportional zu $e^{-\frac{\Delta E}{kT}}$ ansteigt, wobei k die Boltzmann-Konstante ist. Daher bleiben MD-Simulationen bei physiologischer Temperatur leicht in lokalen Minima stecken und ergeben somit kein korrektes Sampling der Energy Landscape innert nützlicher Rechenzeit. In dieser Dissertation wird die *Replica-exchange Methode* (REM) benutzt, um die Faltung und Entfaltung des Beta3s-Peptides bei Temperaturen zwischen $275K$ bis $465K$ zu simulieren. Ausgiebige Vergleiche mit Simulationen bei konstanter Temperatur zeigen, dass REM ein korrektes Sampling der Energy Landscape bei physiologischer Temperatur ermöglicht, was mit herkömmlichen MD-Simulationen nicht möglich ist. Allerdings hat diese Methode den Nachteil, dass sie keine kinetischen Untersuchungen erlaubt. REM-Simulationen wurden des weiteren in einer Studie eingesetzt, die sich mit dem frühen Stadium der Peptid-Aggregation eines amyloidogenetischen Peptides des Prionen-Proteins Sup35 befasst.

Contents

Summary **I**

Zusammenfassung **III**

Contents **V**

1 Energy landscapes **1**
1.1 Introduction 1
1.2 The free-energy perspective 3
1.3 Energy landscapes as complex networks 7
Bibliography 11

2 The protein folding network (JMB (2004) 342, 299) **12**

3 Local modularity measure for network clusterizations (PRE (2005) 72, 056107) **21**

4 Estimation of protein folding probability from equilibrium simulations (JCP (2005), 122, 184901) **26**

5 $\Phi-$value analysis by molecular dynamics simulations of reversible folding (PNAS (2005), 102, 628) **32**

6 Replica exchange molecular dynamics simulations of reversible folding (JCP (2003) 119, 4035) **39**

7 Replica exchange molecular dynamics simulations of amyloid peptide aggregation (JCP (2004), 121, 10748) **48**

Conclusions **58**
Bibliography 60

List of figures **61**

Chapter 1

Energy landscapes

1.1 Introduction

More than 30 years ago, Goldstein articulated a topographic viewpoint of condensed phases [1] that has come to be known as *the energy landscape paradigm* [2]. His seminal ideas have since been applied to protein folding [3, 4, 5, 6, 7], the mechanical properties of glasses [8, 9] and the dynamics of supercooled liquids [10, 11]. Energy landscape is the name generally given to the potential energy function of an $N-$body system $\phi(r1, \ldots, r_N)$, where the vectors r_i comprise position, orientation and vibration coordinates. Figure 1.1 is a schematic illustration of an energy landscape. In condensed phases, whether liquid or solid, every molecule experiences simultaneous interactions with numerous neighbors. Under these conditions it is convenient to consider the full $N-$body $\phi-$function. The landscape is a multidimensional surface. For the simplest case of N structureless particles possessing no internal orientational and vibrational degrees of freedom, the landscape is a $(3N+1)-$dimensional object. The quantities of interest are the number and the nature of the potential energy minima (also called inherent structures) and the barriers between them.

Energy landscape theory provides a framework for the description of the kinetics and thermodynamics of condensed phases. In the past years, it was extensively applied to the analysis of protein folding. Proteins are essential macromolecules for life and are responsible for most cellular functions. However, the molecular processes by which proteins reach their functional structure are not fully understood [13]. Within the energy landscape framework, protein folding is envisioned to proceed along a moderately rough funnel-shaped surface [7]. The overall shape of the landscape arises from a strong energetic

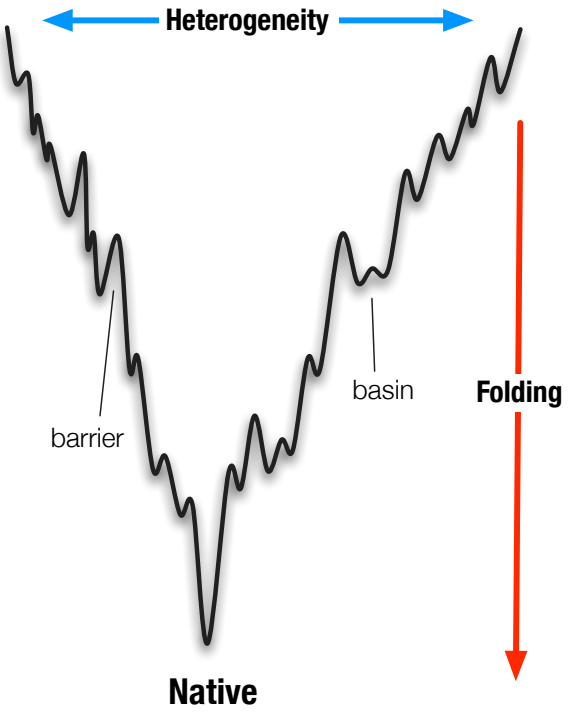

Fig. 1.1: **Funneled energy landscape for folding.** The overall shape of the surface is funnel-like with an energetic bias towards the native state. However, many local minima can arise as a consequence of competing chain interactions and show up as basins of attraction. Kinetics and thermodynamics of folding are strongly influenced by the presence of such heterogeneous non-native basins [12].

driving force to the native global minimum. This energetic bias is necessary to overcome the conformational search problem associated with finding the native state of the protein within a biologically reasonable time frame[1] [7, 14]. The roughness of the surface corresponds to local energy minima arising from the many competing interactions that are possible between the residues. Energetic traps are sequence-related traps and arise when non-native but stabilizing contacts form as the chain folds. For this reason, even if there is a bias towards the native state the energy landscape for folding is rugged. The number and the depth of such energetic traps influence both the thermodynamic and kinetic

[1] In contrast with the astronomical amount of time needed by a random search in the configuration space of the protein (Levinthal paradox).

aspects of folding.

Molecular dynamics simulations (MD) are a very useful tool to study at an atomic detail the dynamics of proteins [15, 16]. In the past years many improvements, both computational and theoretical, have been done to increase computers' speed, accuracy of the models and reliability of the simulations. However, even for a small protein it is currently not yet feasible to simulate reversible folding with a high-resolution approach, e.g., MD simulations with an all-atom model. Despite their limitations, computer simulations are an important theoretical tool for the investigation of the energy landscapes governing protein folding.

The thesis is organized as follows. Chapter 1 continues with a general description of the energy landscape paradigm applied to protein folding. The limits of the approach are presented and a novel framework based on complex networks is introduced. Chapter 2 presents the network approach for the analysis of the beta3s peptide and a random heteropolymer. In chapter 3 a novel measure for the quality of network clusterizations is introduced. Chapter 4 presents a statistical approximation for the protein folding probability. The goal of this chapter is to find a fast and general way to estimate the transition state ensemble out of a set of MD folding simulations. In Chapter 5, the approach is applied to the computation of the $\Phi-$values for the beta3s peptide. Chapters 6 and 7 are dedicated to the problem of limited sampling in MD simulations. The replica exchange method is applied in chapter 6 to the study of folding, while in chapter 7 the method is used for the analysis of the early stages of peptide aggregation. The conclusive chapter is dedicated to my final remarks and future lines of research.

1.2 The free-energy perspective

The kinetics and thermodynamics of a system at temperature T are not only governed by the energetics, i.e. the potential energy $\mathcal{U}$ (internal energy or *enthalpy*), but also by the *entropy* $\mathcal{S}$ that, in the case of a protein, is conformational entropy. The free-energy of a protein in configuration i at temperature T is written as

$$\mathcal{F}_i = \mathcal{U}_i - T\mathcal{S}_i \tag{1.1}$$

where $\mathcal{U}_i$ and $\mathcal{S}_i$ are the internal energy and the conformational entropy of configuration i, respectively. In simple words, the conformational entropy of a protein reflects in how many microscopic ways (arrangement of atoms), at a given temperature T, it is possible to realize a given *configuration* or *state*

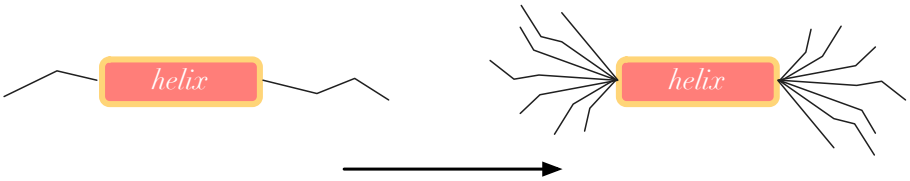

Fig. 1.2: **Entropic stabilization.** Despite an unfavorable internal energy, partially structured conformations can be stabilized entropically as in the case of some helical conformations of beta3s.

of the protein. As an example consider the partially helical configuration observed in MD simulations of beta3s[2] (Figure 1.4, top left). This configuration, let's call it HH, is α helical at the central residues and mainly unstructured elsewhere. Despite the unfavorable internal energy compared to more compact configurations, HH is a free-energy basin in the denatured state of beta3s. This conformation is stabilized entropically because there is a large number of possible arrangements of the tails compatible to it (Figure 1.2). It is the overall balance between the enthalpic and entropic contributions that determines at a given T the stability, i.e. the free-energy, of a state. A consequence of these considerations is that the states of a protein can be automatically defined as the minima of the free-energy landscape.

Order parameters. A useful way to investigate and display the free-energy landscape is to study it as a function of one or more *order parameters*, i.e., suitably chosen macroscopic quantities that distinguish the different states of the protein [21]. For example, it is common in the study of protein folding to use the number of native contacts Q [18]. Q is a good *order parameter* in the sense that it distinguishes the unfolded and folded states: unfolded states typically have small Q, while by definition $Q = Q_{max}$ in the native state. The free-energy of a protein as a function of Q can be written as

$$\mathcal{F}(Q) = \mathcal{U}(Q) - T\mathcal{S}(Q) \tag{1.2}$$

where $\mathcal{F}(Q)$, $\mathcal{U}(Q)$ and $\mathcal{S}(Q)$ are the average free-energy, potential energy and configurational entropy for a configuration with Q native contacts, respectively.

[2]Beta3s a designed peptide whose solution conformation has been studied by NMR [17]. The NMR data indicate that Beta3s forms a monomeric three-stranded antiparallel β-sheet conformation with turns at Gly_6-Ser_7 and $Gly_{14}-Ser_{15}$. In previous studies have shown that it is possible to simulate the reversible folding of beta3s at relatively high temperature values (330-360 K) [18, 19] using an implicit model of the solvent based on the accessible surface area [20].

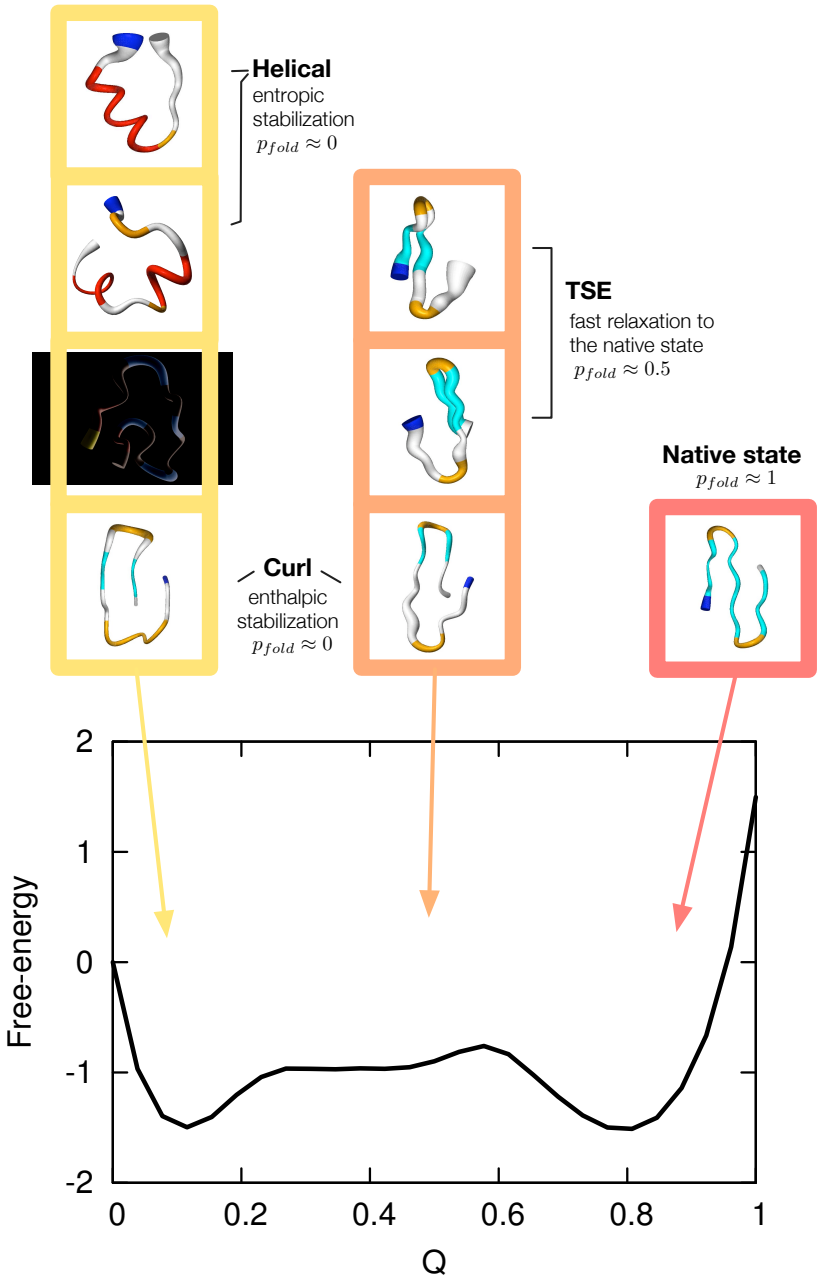

Fig. 1.3: **Free-energy projections on order parameters.** In the case of beta3s, the fraction of native contacts doesn't necessarily identify structurally and kinetically homogeneous conformations. In the first, second and third column, conformations with $\sim 30\%$, $\sim 60\%$ and $\sim 90\%$ (native state) of native contacts Q are shown, respectively. The projected free-energy shows no evidence of the structurally and kinetically heterogeneity found in the denatured state of beta3s (see text).

Free-energy projections on order parameters are usually used for the understanding and the analysis of many aspects of protein folding. *States* are associated with local free energy minima of the projected landscape. The deepness of the minima is considered proportional to the stability of the states associated to them and the barriers between different minima indicate activation energies between states. In many cases, this approach reveals a surprisingly *simple* two-state picture for protein folding (Figure 1.3, bottom). Order parameters are also used as reaction coordinates to monitor the dynamics of the protein [21]. Using free-energy projections for the study of the kinetics requires knowledge of a good reaction coordinate for folding, which is a difficult and unresolved problem [22]. Given the complexity of protein folding and the large number of degrees of freedom involved, reaction coordinates are often arbitrary chosen and can miss essential aspects of the process [23, 24].

Some hidden assumptions are made to justify the approach just described above:

(i) The order parameter(s) used are sufficient to distinguish the various states of the system.

(ii) Within a minimum, conformations can interconvert rapidly.

A first consequence of assumption (i) is that every value of the order parameter (or the combination of different order parameters) identifies only one state of the system. Assumption (ii), stated in a different way, says that all the conformations in a state are kinetically homogeneous. This means for example that, if Q is used as order parameter, all the conformations with half of native contacts should take similar times to go to the native state. We will see below that this is not generally true and the case of beta3s is very instructive in this sense. In Figure 1.3, some representative conformations of beta3s with $\approx 30\%$, $\approx 60\%$ and $\approx 90\%$ of native contacts are shown, from left to right, respectively [23]. Q identifies uniquely one state only when almost all the native contacts are formed, i.e., the native state. For $Q < 70-80\%$ many heteregenous conformations can have the same number of native contacts [23]. Most of the times, these conformations are structurally and kinetically heterogeneous. To verify this last point, the cluster-p_{fold} [25], introduced in Chapter 4, was computed in order to estimate the kinetic distance from the native state. Conformations with half of the native contacts (central column in Figure 1.3) can either have $p_{fold} \approx 0$ (curl-like structure), i.e., conformations that are far away with respect to the native state, or $p_{fold} \approx 0.5$, i.e. conformations on top of the folding barrier. For values of $Q \approx 0$ folding times can differ as much as one order of magnitude.

Of course it can be objected that, in order to optimally describe the thermodynamics and the kinetics of a peptide or a protein, suitable combinations of order parameters can always be found [26]. Even if this possibility exists, it is either very difficult to find and/or very specific for the system under study. A large part of my thesis work was dedicated to improve the description of energy landscapes.

1.3 Energy landscapes as complex networks

In the last five years, many complex systems, like the Internet [27], social interactions [28], metabolic pathways [29] and protein structures [30] have been modeled as networks. Networks have shown to be a comprehensive and universal approach for the description of complex systems [31]. Given this universality, we applied the network framework to the description of the free-energy landscape of proteins [23]. This approach is proposed in order to overcome the limitations presented in the last section concerning projected free-energy surfaces. A network description of the energy landscape is able to keep the *structural* and *kinetic* properties of the conformations involved into the folding process. At the same time the network approach gives a synthetic 2-dimensional view of the multi-dimensional free-energy surface.

For this purpose, a discretization of the conformational space of the protein is needed [32, 33]. From here on, the term conformation is used for a set of snapshots grouped together by means of their structural similarity[3]. Given this definition, conformations and the transitions between them observed in a MD simulation are nodes and links of the network, respectively.

The first application of the method was to the beta3s peptide [23]. Figure 1.4 shows the free-energy landscape of the beta3s peptide, represented as a network. Red nodes belong to the native basin while all the other nodes build up the denatured state. Simulations were run at the melting temperature, so the native basin contains $\approx 50\%$ of the total number of snapshots sampled during the MD simulations. The picture of the free-energy landscape obtained within the network framework is pretty different from the traditional two-state view of the folding of beta3s. The landscape is characterized by one well pronounced free-energy basin corresponding to the native state and a large and heterogeneous denatured state defined by non-native metastable free-energy

[3]Many methods have been proposed and all of them give similar results provided that snapshots representing a given conformation are structurally and kinetically homogeneous [25].

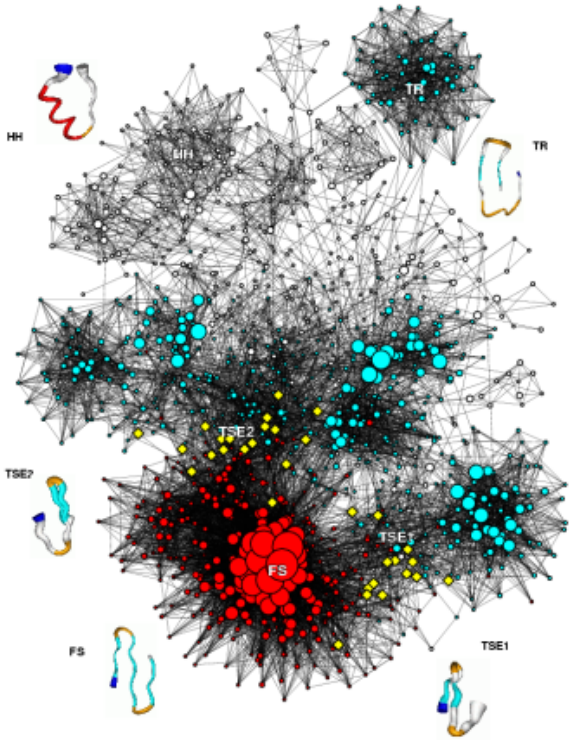

Fig. 1.4: **Beta3s free-energy landscape network.** Nodes and links are the conformations and the transitions between them, respectively. Red nodes represents the native basin. Node size is inversely proportional to the free-energy of the conformation.

basins. Some of those basins are stabilized entropically, others enthalpically. The helical basin belongs to the former, because only a small fraction of the chain is structured (Figure 1.4, top left). An example of enthalpic stabilization is the curl-like basin shown in figure 1.4, which is extremely rigid as a consequence of the multiple interactions of the terminal residues inside the curl. For beta3s, barriers between the metastable basins appear to be higher than the folding barrier. In contrast to the usual picture of protein folding [24], no fast inter-conversion in the denatured state is observed. The presence of metastable non-native conformations, that is also consistent with a recent experimental work [34], points out important questions on the nature and organization of the denatured state as well as the origin of the folding barrier. From these results it appears that, in the case of the beta3s peptide, the simple funnel picture for folding needs to be revisited.

Bibliography

[1] M. Goldstein. Viscous liquids and glass transition - a potential energy barrier picture. *Journal of Chemical Physics*, 51:3728, 1969.

[2] F.H. Stillinger. A topographic view of supercooled liquids and glass-formation. *Science*, 267:1935–1939, Mar 1995.

[3] H. Frauenfelder, S.G. Sligar, and P.G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, Dec 1991.

[4] J.G. Saven, J. Wang, and P.G. Wolynes. Kinetics of protein-folding - the dynamics of globally connected rough energy landscapes with biases. *Journal of Chemical Physics*, 101:11037–11043, Dec 1994.

[5] V.I. Abkevich, A.M. Gutin, and E.I. Shakhnovich. Free-energy landscape for protein-folding kinetics - intermediates, traps, and multiple pathways in theory and lattice model simulations. *Journal of Chemical Physics*, 101:6052–6062, Oct 1994.

[6] J. Wang, J. Onuchic, and P. G. Wolynes. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Physical Review Letters*, 76:4861–4864, Jun 1996.

[7] K.A. Dill and H.S. Chan. From levinthal to pathways to funnels. *Nature Structural Biology*, 4:10–19, Jan 1997.

[8] D.L. Malandro and D.J. Lacks. Volume dependence of potential energy landscapes in glasses. *Journal of Chemical Physics*, 107:5804–5810, Oct 1997.

[9] D.J. Lacks. Localized mechanical instabilities and structural transformations in silica glass under high pressure. *Physical Review Letters*, 80:5385–5388, Jun 1998.

[10] S. Sastry, P.G. Debenedetti, and F.H. Stillinger. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature*, 393:554–557, Jun 1998.

[11] M. Schulz. Energy landscape, minimum points, and non-arrhenius behavior of supercooled liquids. *Physical Review B*, 57:11319–11333, May 1998.

[12] S.F. Chekmarev, S.V. Krivov, and M. Karplus. Folding time distributions as an approach to protein folding kinetics. *Journal of Physical Chemistry B*, 109:5312–5330, 2005.

[13] V. Daggett and A.R. Fersht. Is there a unifying mechanism for protein folding? *Trends In Biochemical Sciences*, 28:18–25, Jan 2003.

[14] M. Karplus. The levinthal paradox: yesterday and today. *Folding & Design*, 2:S69–S75, 1997.

[15] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[16] M. Karplus and J. Kuriyan. Chemical theory and computation special feature: Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA*, 102:6679–6685, 2005.

[17] E. De Alba, J. Santoro, M. Rico, and M.A. Jimenez. De novo design of a monomeric three-stranded antiparallel beta-sheet. *Protein Science*, 8:854–865, Apr 1999.

[18] P. Ferrara and A. Caflisch. Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc. Natl. Acad. Sci. USA*, 97:10780–10785, Sep 2000.

[19] A. Cavalli, P. Ferrara, and A. Caflisch. Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Structure Function And Genetics*, 47:305–314, May 2002.

[20] P. Ferrara, J. Apostolakis, and A. Caflisch. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure Function And Genetics*, 46:24–33, Jan 2002.

[21] R. Du, V.S. Pande, A.Y. Grosberg, T. Tanaka, and E.S. Shakhnovich. On the transition coordinate for protein folding. *Journal of Chemical Physics*, 108:334–350, Jan 1998.

[22] V.S. Pande, A.Y. Grosberg, T. Tanaka, and D.S. Rokhsar. Pathways for protein folding: is a new view needed? *Current Opinion In Structural Biology*, 8:68–79, Feb 1998.

[23] F. Rao and A. Caflisch. The protein folding network. *Journal of Molecular Biology*, 342:299–306, 2004.

[24] S.V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA*, 101:14766–14770, Oct 2004.

[25] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch. Estimation of protein folding probability from equilibrium simulations. *Journal of Chemical Physics*, 122:184901, 2005.

[26] R.B. Best and G. Hummer. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA*, Early edition, 2005.

[27] R. Albert, H. Jeong, and A.L. Barabasi. Internet - diameter of the world-wide web. *Nature*, 401:130–131, Sep 1999.

[28] M.E.J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98:404–409, Jan 2001.

[29] H. Jeong, B. Tombor, R. Albert, Z.N. Oltval, and A.L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, Oct 2000.

[30] L.H. Greene and V.A. Higman. Uncovering network systems within protein structures. *Journal of Molecular Biology*, 334:781–791, Dec 2003.

[31] M.E.J. Newman. The structure and function of complex networks. *Siam Review*, 45:167–256, Jun 2003.

[32] P. Carter, C.A. Andersen, and B Rost. Dsspcont: continuous secondary structure assignments for proteins. *Nucleic Acids Research*, 31:3293–3295, Jul 2003.

[33] J.A. Hartigan. *Clustering algorithms*. Wiley, New York, 1975.

[34] Y.F. Tang, D.J. Rigotti, R. Fairman, and D.P. Raleigh. Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: The villin headpiece subdomain. *Biochemistry*, 43:3264–3272, Mar 2004.

Chapter 2

The protein folding network

(JMB (2004) 342, 299)

doi:10.1016/j.jmb.2004.06.063 J. Mol. Biol. (2004) 342, 299–306

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Protein Folding Network

Francesco Rao and Amedeo Caflisch*

Department of Biochemistry
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich, Switzerland

The conformation space of a 20 residue antiparallel β-sheet peptide, sampled by molecular dynamics simulations, is mapped to a network. Snapshots saved along the trajectory are grouped according to secondary structure into nodes of the network and the transitions between them are links. The conformation space network describes the significant free energy minima and their dynamic connectivity without requiring arbitrarily chosen reaction coordinates. As previously found for the Internet and the World-Wide Web as well as for social and biological networks, the conformation space network is scale-free and contains highly connected hubs like the native state which is the most populated free energy basin. Furthermore, the native basin exhibits a hierarchical organization, which is not found for a random heteropolymer lacking a predominant free-energy minimum. The network topology is used to identify conformations in the folding transition state (TS) ensemble, and provides a basis for understanding the heterogeneity of the TS and denatured state ensemble as well as the existence of multiple pathways.

Keywords: complex networks; protein folding; energy landscape; transition state; denatured state ensemble

*Corresponding author

Proteins are complex macromolecules with many degrees of freedom. To fulfil their function they have to fold to a unique three-dimensional structure (native state). Protein folding is a complex process governed by non-covalent interactions involving the entire molecule. Spontaneous folding in a time-range of microseconds to seconds[1] can be reconciled with the large amount of conformers by using energy landscape analysis.[2–4] The main difficulty of this analysis is that the free energy has to be projected on arbitrarily chosen reaction coordinates (or order parameters). In many cases, a simplified representation of the free-energy landscape is obtained where important information on the non-native conformation ensemble and the folding TS ensemble are hidden. Moreover, the possible transitions between free-energy minima cannot be displayed in such projections, which hinders the study of pathways and folding intermediates. The characterization of the free-energy minima and the connectivity among them, i.e. possible transitions between minima, for peptides and proteins is still a challenging problem despite the fact that several elegant approaches have been proposed.[5–7]

In the last five years, many complex systems, like the World-Wide Web, metabolic pathways, and protein structures have been modeled as networks.[8–11] Intriguingly, common topological properties have emerged from their organization.[12] The conformation space of a short two-dimensional lattice polymer chain has been mapped to a network where a link between two nodes indicates the interconversion in a single Monte Carlo move of the chain.[13] A description of the potential energy landscape without the use of any projection has been given in terms of networks for a Lennard–Jones cluster of atoms.[14]

Here, we use complex network analysis[12] to study the conformation space and folding of beta3s, a designed 20 residue sequence whose solution conformation has been investigated by NMR spectroscopy.[15] The NMR data indicate that beta3s in aqueous solution forms a monomeric (up to more than 1 mM concentration) triple-stranded antiparallel β-sheet (Figure 1, bottom), in equilibrium with the denatured state.[15] We have previously shown that in implicit solvent[16] molecular dynamics simulations beta3s folds reversibly to the NMR solution conformation, irrespective of the starting conformation.[17,18] We consider

Abbreviations used: RMSD, root-mean-square deviations; TS, transition state; TR, trap; FS, folded state.

E-mail address of the corresponding author: caflisch@bioc.unizh.ch

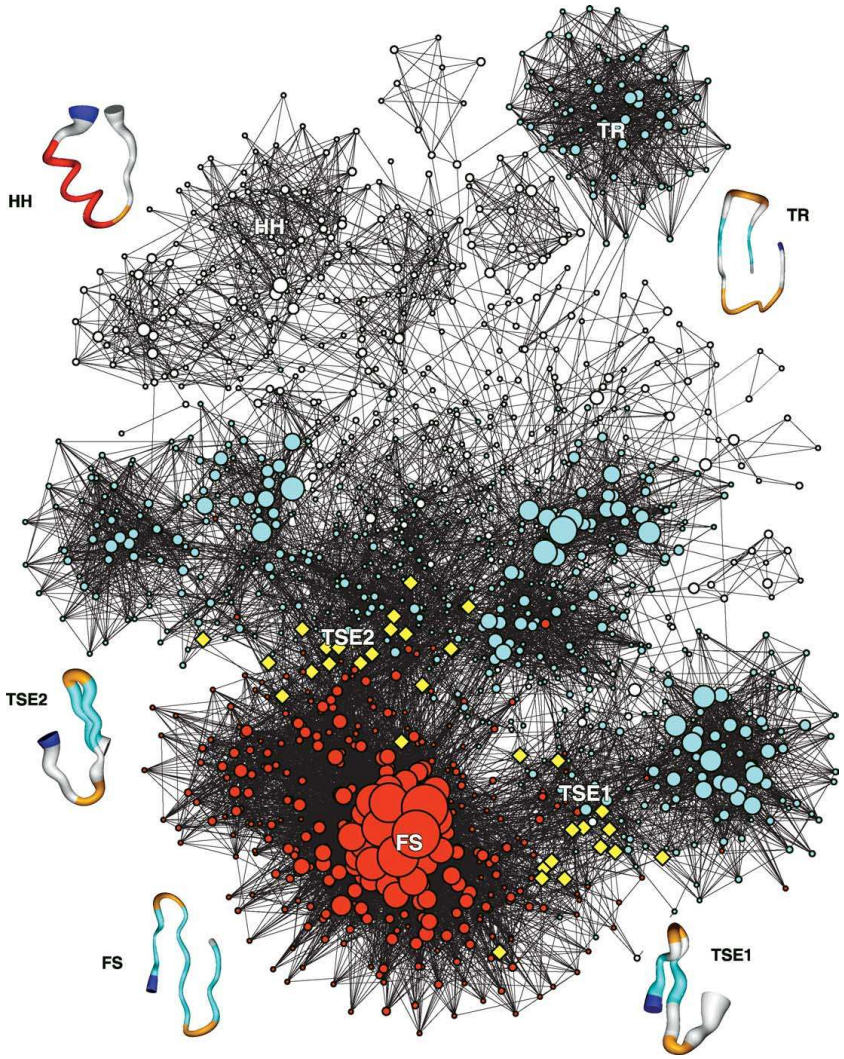

Figure 1. The beta3s conformation space network. The size and color coding of the nodes reflect the statistical weight w and average neighbor connectivity k_{nn} respectively. White, cyan, and red nodes have $k_{nn}<30$, $30\leq k_{nn}\leq 70$, and $k_{nn}>70$, respectively. Representative conformations are shown by a pipe colored according to secondary structure: white stands for coil, red for α-helix, orange for bend, cyan for strand and the N terminus is in blue. The variable radius of the pipe reflects structural variability within snapshots in a conformation. The yellow diamonds are folding TS conformations (TSE1, TSE2, see the text for details) characterized by a connectivity/weight ratio $k/2\bar{w}>0.3$, a clustering coefficient $C<0.3$, and $60<k_{nn}<80$. This Figure was made using visone (www.visone.de) and MOLMOL[40] visualization tools.

conformations sampled by molecular dynamics simulations and the transitions between them as the network nodes and links, respectively. The network analysis allows us to identify the topological properties that are common to both beta3s, which folds to a unique three-dimensional structure,[15,19] and a random heteropolymer which lacks a single preferential conformation like the native state despite the fact that it has the same residue composition as beta3s. These properties include the presence of several free-energy minima and highly connected conformations (hubs). On the other hand, a hierarchical modularity[20] in the proximity of the native state is peculiar of a folding sequence.

Model and Methods

Molecular dynamics simulations

The simulations and part of the analysis of the trajectories were performed with the program CHARMM.[21] beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field.[21]) A mean field approximation based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute.[16] The two parameters of the solvation model were optimized without using beta3s. The same force field and implicit solvent model have been used recently in molecular dynamics simulations of the early steps of ordered aggregation,[22] and folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues,[16,17,23] as well as small proteins of about 60 residues.[24,25] Despite the absence of collisions with water molecules, in the simulations with implicit solvent the separation of time-scales is comparable with that observed experimentally. Helices fold in about 1 ns,[26] β-hairpins in about 10 ns[26] and triple-stranded β-sheets in about 100 ns,[18] while the experimental values are ∼0.1 μs,[27] ∼1 μs[27] and ∼10 μs,[15] respectively. Recently, four molecular dynamics simulations of beta3s were performed at 330 K for a total simulation time of 12.6 μs.[19] There are 72 folding events and 73 unfolding events, and the average time required to go from the denatured state to the folded conformation is 83 ns. The 12.6 μs of simulation length is about two orders of magnitude longer than the average folding or unfolding time, which are similar because at 330 K the native and denatured states are almost equally populated.[19] For the network analysis the first 0.65 μs of each of the four simulations were neglected so that along the 10 μs of simulations there are a total of 5×10^5 snapshots because coordinates were saved every 20 ps. The sequence of the random heteropolymer is a randomly scrambled version of the beta3s sequence with the same residue composition. It was simulated for 2 μs and 10^5 snapshots were saved. The conditions for the molecular dynamics simulations, i.e. force field, solvation model, temperature, and time interval between saved snapshots were the same for both peptides.

Construction of the protein folding network

To define the nodes and links of the network the secondary structure was calculated[28] for each snapshot (Cartesian coordinates of the atomic nuclei) saved along the molecular dynamics trajectory. A "conformation" is a single string of secondary structure,[28] e.g., the most populated conformation for beta3s (FS in Figure 1) is:

```
-EEEESSEEEEEES SEEEE-
```

There are eight possible "letters" in the secondary structure "alphabet":

"H", "G", "I", "E" "B", "T", "S", and "-",

standing for α-helix, 3_{10} helix, π-helix, extended, isolated β-bridge, hydrogen bonded turn, bend, and unstructured, respectively. Since the N and C-terminal residues are always assigned an "-"[28] a 20 residue peptide can, in principle, assume $8^{18}\approx10^{16}$ conformations. Conformations are nodes of the network and the transitions between them are links. A weight $\bar{w}$ is assigned to each node to take into account the free-energy of each conformation and is equal to the number of snapshots with a given secondary structure string. The statistical weight w of a node is equal to $w=\bar{w}/N$, where N is the total number of snapshots in the simulation (N is equal to 5×10^5 and 10^5 for beta3s and the random heteropolymer, respectively). Considering all the conformations visited during a microsecond-scale simulation can yield to a computationally intractable network size. For this reason we used for the network analysis the 1287 conformations of beta3s with significant weight ($\bar{w}\geq20$ per conformation). Two nodes are connected by an undirected link (and called neighbors) if they either include a pair of snapshots that are visited within 20 ps or they are separated by one or more conformations with less than 20 snapshots each. For the 2 μs of the random heteropolymer, a threshold of $\bar{w}\geq4$ was used, so that $w\geq4\times10^{-5}$ as in the beta3s network. The choice of a threshold value is somewhat arbitrary but the network properties are robust for a large range of threshold values (see Supplementary Material).

The properties of the network are robust also with respect to the length of the simulation time and the definition of the nodes. The topological properties are independent from simulation lengths if one considers more than 2 μs. The correlation between statistical weight and connectivity, as well as power-law behavior of the connectivity distribution, and $1/k$ behavior of the clustering coefficient distribution (see below) are essentially identical after 2 μs, 4 μs, and 10 μs. As an example, the exponent of the power-law is 2.0 for the beta3s networks based on 2 μs, 4 μs and 10 μs of simulation time. Defining nodes by grouping snapshots according to root-mean-square deviations (RMSD) in coordinates of C^α–C^β atoms yields the same overall properties, i.e. power-law distribution of the links (with a scaling factor γ of 2.2) and $1/k$ tail of the clustering distribution. Grouping snapshots according to secondary structure motifs does not require the use of an arbitrarily chosen RMSD cutoff, and is able to capture the fluctuations of partially structured conformations.[28]

Evaluation of P_{fold}

The TS ensemble can be defined as the set of structures which have the same probability of folding (P_{fold}) or unfolding in trajectories started with varying initial conditions.[29] For each putative TS conformation, the probability to fold before unfolding was calculated by 100 very short

Table 1. Energetic comparison of folded and denatured state

	$\langle E\rangle^{a}$	$\langle\Delta\mathcal{F}\rangle^{b}$
Folded state (FS)		
-EEEESSEEEEEESSEEEE-	−7.6	0
-EEE-STTEEEEESEEEE-	−8.6	0.1
-EEEESSEEEEE-STTEEE-	−8.4	0.5
-EEE-STTEEEE-STTEEE-	−9.2	0.7
Helical conformations (HH)		
---HHHHHHHHHS------	0.9	3.1
-HHHHHHHHHHS------	−1.9	3.3
---HHHHHHHHHTT-----	0.7	3.5
---HHHHHHHHH-------	0.5	3.7
-HHHHHHHHHHTT-----	−0.8	3.7
--TT--HHHHHHHHHHH-	−0.8	3.8
Curl-like trap (TR)		
---SSGGG-EEE-STTTEE-	−7.8	3.4
---SSSS--EEE-STTTEE-	−7.0	3.5
---S-GGG-EEE-STTTEE-	−9.3	3.7
---SSGGG-EEE-SGGGEE-	−9.6	3.7
---SSTTT-EEE-STTTEE-	−8.4	3.7

The free-energy of conformation i is $\mathcal{F}_i = -k_B T\, log(w_i)$, where w_i is the probability along the trajectory to find the peptide in the conformation i.

[a] Average effective energy.

[b] Free-energy relative to the most populated conformation. All values are in kcal/mol. The conformational entropy of the peptide is equal to $(\langle\mathcal{E}\rangle - \mathcal{F})/T$. Note that the curl-like traps are entropically penalized with respect to the native state.

trajectories at 330 K started from ten snapshots within a node. The only difference between the ten runs was the seed for the random number generator used for the initial assignment of the atomic velocities. A trajectory was considered to lead to folding (unfolding) if it visits first structures with a fraction of native contacts Q>22/26 (Q<4/26).[17] The 33,381 snapshots with Q>22/26 have a distribution of the pairwise C^{α} RMSD peaked at 1.1 Å (see Supplementary Material).

Results and Discussion

To study the conformation space network of polypeptides we concentrate on the analysis of topology, i.e. on the study of the connectivity between different conformations, leaving for a later study the analysis of transition rates. We have investigated the network topologies of several peptides but, here, we focus on beta3s and the random scrambled version of it. Additional details can be found in the Supplementary Material, where the network properties of another structured peptide and a glycine homopolymer are presented.

Conformation space network of a structured peptide

The conformation space network and relevant structures of beta3s are shown in Figure 1. The group of nodes at the bottom of Figure 1 (red nodes) represents the native state basin (FS). The native basin is connected to a wide region of nodes with significant native content (cyan circles in the middle of Figure 1). Although many heterogeneous routes can be taken to reach the folded state (in agreement with lattice simulations),[30,31] most of the folding events have common structural features that define two average folding pathways. The less frequented average pathway[18] (see the density of transitions in Figure 1, bottom right) consists of conformations that have the N-terminal hairpin formed while the C-terminal strand is mostly unstructured with non-native hydrogen bonds at the turn (TSE1 in Figure 1). The second and most frequented average pathway includes conformations with a well formed C-terminal hairpin while the N-terminal strand is disordered (TSE2 in Figure 1), namely it can be out-of-register or mostly unstructured. It is interesting to note that the same two folding pathways were observed experimentally for a 24 residue peptide with the same folded state as beta3s.[32] Furthermore, multiple folding pathways have recently been detected by kinetic analysis of a β-sandwich protein.[33]

The denatured state ensemble is very heterogeneous and includes high-enthalpy, high-entropy conformations (e.g. the partially helical conformations, denoted HH in Figure 1) but also low-enthalpy, low-entropy conformations (e.g., the curl-like trap, TR). The former are loosely linked clusters of conformations with similar secondary structure (see Table 1) which are characterized by an unfavorable effective energy (sum of peptide potential energy and solvation energy) and fluctuating unstructured residues (e.g. the terminal of the helix shown on top left of Figure 1). On the contrary, low-enthalpy, low-entropy traps form tightly linked clusters with almost identical secondary and tertiary structure, favorable effective energy (similar to the one of the native structure, see Table 1) and no fluctuating residues (e.g. Figure 1, top right). Taken together, these results indicate that FS is entropically favored over low-enthalpy conformations like TR, i.e. FS has more flexibility than TR. A possible

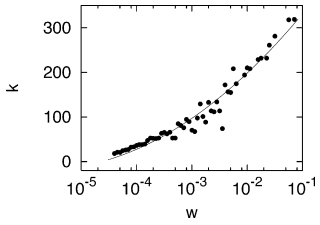

Figure 2. Correlation between the statistical weight w and the connectivity k for beta3s. The connectivity can be fitted to $\log^2(w)$ (with a correlation coefficient of 0.88, continuous line) indicating a deviation from a purely diffusive dynamics where $k \sim w$. The correlation and the fit are calculated over all nodes of the network but in the Figure logarithmic binning is applied to reduce noise.

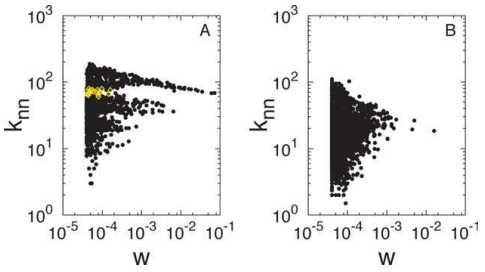

Figure 3. Average neighbor connectivity k_{nn} plotted as a function of the statistical weight for the 1287 nodes of beta3s (A) and for the 2658 nodes of the random heteropolymer (B). k_{nn} of node i is the average number of links of the neighbors of node i. The yellow diamonds are folding TS conformations (see also Figure 1 and the text) characterized by a connectivity/weight ratio $k/2\bar{w} > 0.3$, a clustering coefficient $C < 0.3$, and $60 < k_{nn} < 80$.

explanation is that the C-terminal carboxy group is involved in four hydrogen bonds in TR (with the backbone NH groups of residues 4–7), whereas both termini undergo rather large fluctuations in FS. In addition, a more favorable van der Waals energy in TR is consistent with a denser packing in TR than in FS. Entropically favored structures (like FS) are destabilized by lowering the temperature. Hence, there should be a temperature (not accessible to conventional MD simulations) where the system becomes frustrated and a glass-like scenario emerges.

Note that the network description of non-native conformations is more detailed than the one obtained by projecting the free energy surface on progress variables (e.g. based on fraction of native contacts). In such projections, for low values of the fraction of native contacts structures as diverse as helices and the curl-like conformations mentioned above are not distinguished. Even the ensemble with half of the native contacts is heterogeneous and hard to classify. Using as reaction coordinate the RMSD (with respect to a given structure) or the radius of gyration is even less selective. Only when a clever combination of variables is used is it possible to have a more detailed description of the free-energy landscape. The network description of the conformation space gives a synthetic and systematic view of all the possible conformations accessed by the system and their transitions. By considering the statistical weight of the nodes a thermodynamical description of the system is obtained.

The high correlation between the statistical weight of a node and its number of links (Figure 2) shows that the most connected nodes are also low-lying minima on the free-energy landscape. This indicates that the conformation space network describes the significant free energy minima and their dynamic connectivity, without projection, where highly populated nodes are minima of free-energy and the set of nodes densely connected to them make up the basins of such minima. The connectivity can be fitted to $\log^2(w)$, which indicates that the dynamics is not diffusive (see Figure 2).

Folding and network topology

The average neighbor connectivity k_{nn} of beta3s (Figure 3A), i.e. the average number of links of the neighbors of a given node, is rather heterogeneous, highlighting the presence of different connection rules in different regions of the network. This is not the case for the random heteropolymer (Figure 3B), whose basins have organization and statistical

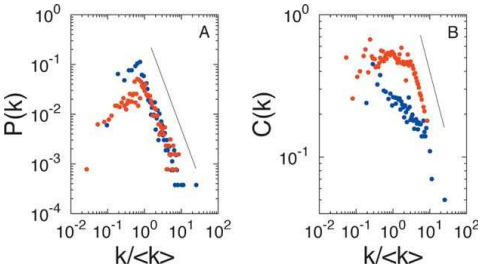

Figure 4. Topological properties of conformation space networks. Red and blue data points are plotted for beta3s and a random heteropolymer, respectively. For a direct comparison, the connectivity k is normalized by the average connectivity $\langle k \rangle$ of each network. Logarithmic binning is applied to reduce noise. A, The connectivity distribution $P(k)$ is the probability that a node (conformation) has k links (neighbor conformations). The straight line corresponds to a power-law fit $y = x^{-\gamma}$ on the tail of the distribution with $\gamma = 2.0$. B, The clustering coefficient C describes the cliques of a node. For node i it is defined as $C_i = 2n_i/k_i(k_i - 1)$, where k_i is the number of neighbors of node i and n_i is the total number of connections between them. Values of C are averaged over the nodes with k links. The straight line corresponds to a power-law fit $y = x^{-1}$ on the tail of the distribution of beta3s.

weight similar among each other as previously found for most homopolymers.[10] Note that for beta3s the native state is well discriminated by k_{nn} (red nodes in Figure 1 and top band in Figure 3A).

The connectivity distribution of conformation space networks shows a well pronounced power-law tail $P(k) \sim k^{-\gamma}$ with $\gamma = 2.0$ for both beta3s and the random heteropolymer (Figure 4A) as well as another structured peptide[34] and homoglycine, i.e. $(Gly)_{20}$ (see Supplementary Material). The power-law is due to the presence of a few largely connected "hubs" while the majority of the nodes have a relatively small number of links.[35] This behavior has been previously observed for several biological,[8] social[36] and technological networks,[9] which in the literature take the name of scale-free networks. In terms of free energy this means that only a few low lying minima are present but they act as "hubs" with a large number of routes to access them.

The average clustering coefficient C is a measure of the probability that any two neighbors of a node are connected. beta3s and the heteropolymer have C values of 0.49 and 0.28, respectively. These values are one order of magnitude larger than random realizations of the two networks with the same amount of nodes and links. The native basin of beta3s includes the nodes with the largest number of links of the network. These nodes give rise to the $1/k$ tail of the clustering distribution (Figure 4B), i.e. an inherently hierarchical organization[20] of the conformations in the native basin of beta3s. Such organization is not apparent for the non-native region of beta3s and the random heteropolymer. Note that the power-law scaling of the connectivity distribution can be considered as a general property of free-energy landscapes of polypeptides, whereas a hierarchical organization of the nodes reflects a pronounced free-energy basin of attraction (like the native state).

Transition state ensemble

As mentioned above, folding is a complex process with many degrees of freedom involved and it is difficult (or even not possible) to define a single reaction coordinate to monitor folding events.[37,38] Hence, it is very difficult to isolate transition state (TS) conformations from equilibrium sampling. The TS conformations are saddle points, i.e. local maxima with respect to the reaction coordinate for folding and local minima with respect to all other coordinates. For this reason, we identified the nodes with a high connectivity/weight ratio $k_i/2\bar{w} > 0.3$ and low clustering coefficient value C_i as putative TS conformations. The former criterion guarantees that these nodes are accessed and exited, most of the time, by a different route, i.e. they can be directly reached from different conformations of the network space. The low clustering coefficient value guarantees that the neighbors of these conformations are likely to be disconnected. These two conditions are necessary

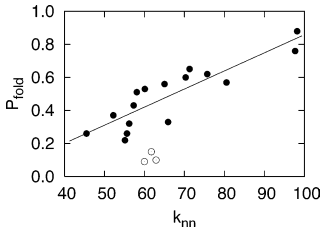

Figure 5. Correlation between P_{fold} and average neighbor connectivity k_{nn}. Three nodes used as a negative control (low connectivity/weight ratio and/or high clustering coefficient but similar fraction of native contacts) are shown with open circles.

but not sufficient because they do not distinguish folding TS conformations from saddle points between unfolded conformations. Since the folding TS conformations are linked to both nodes in the native state (having large number of links) and in the denatured state (small/intermediate number of links), we speculated that folding TS conformations should have values of the average neighbor connectivity k_{nn} within a certain range. For nodes with high connectivity/weight ratio and low clustering coefficient, a remarkable correlation of 0.89 was found between the average neighbor connectivity k_{nn} and P_{fold} (Figure 5), which is the probability of a given conformation to fold before unfolding.[29] A P_{fold} value close to 0.5 is expected for conformations on top of the folding TS barrier[25] and the correlation suggests that network properties can be used to predict folding TS conformations. These are shown in Figures 1 and 3A with yellow diamonds. As discussed above, two main average folding pathways are observed. The less frequent one is characterized by a TS ensemble of conformations with the first hairpin in a native form (residues 1–13) and a bend corresponding to the second native turn (residues 14 and 15). The C-terminal residues form a straight structure with almost no contacts, either native or non-native. The second average pathway shows a TS with the second native hairpin formed (residues 7–20) and a bend corresponding to the first native turn (residues 5 and 6). Such a symmetrical behavior is presumably due to the simplicity and symmetry of the native conformation as well as the symmetry in the sequence (sequence identity of 67% between the two hairpins). The folding TS conformations of beta3s form a heterogeneous ensemble with C^{α} RMSD within contributing structures between 3 Å and 6 Å. In contrast to previous molecular dynamics studies in which progress variables based on fraction of native contacts were used to describe TS conformations,[17,39] the network properties yield a description of the folding TS ensemble (Figure 1) which does not depend on the choice of reaction coordinates. Interestingly, the folding TS conformations of beta3s have about one-half of the

native contacts formed but this is not a sufficient criterion (Table S1 in Supplementary Material). Moreover, there is no correlation between the fraction of native contacts and the probability of folding. As a control, P_{fold} values smaller than 0.15 were obtained for five nodes with an average fraction of native contacts similar to the folding TS conformations but low connectivity/weight ratio and/or high clustering coefficient.

Conclusions

Complex network theory was used to analyze the conformation space of a structured peptide and that of a random heteropolymer of the same residue composition. Four main results have emerged. First, as it was already observed for a variety of networks as diverse as the World-Wide Web and the protein interactions in a cell, the conformation space network of polypeptide chains is a scale-free network (power-law behavior of the degree distribution). Second, the native basin of the structured peptide shows a hierarchical organization of conformations. This organization is not observed for the random heteropolymer which lacks a native state. Third, free energy minima and their connectivity emerge from the network analysis without requiring projections into arbitrarily chosen reaction coordinates. As a consequence, it is found that the denatured state ensemble is very heterogeneous and includes high-entropy, high-enthalpy conformations as well as low-entropy, low-enthalpy traps. Fourth, the network properties were used to identify TS conformations and two main average folding pathways. It was found that the average neighbor connectivity k_{nn} correlates with P_{fold}, the probability of folding. P_{fold} is computationally very expensive to evaluate. Hence, it will be important to generalize this result by analyzing other structured peptides, which is work in progress in our research group. In conclusion, the network analysis seems to be particularly useful to study the conformation space and folding of structured peptides including the otherwise elusive TS ensemble.

Acknowledgements

We thank M. Cecchini, Professor P. De Los Rios, E. Guarnera, Professor M. Karplus, Dr E. Paci, Dr M. Seeber and Dr G. Settanni for interesting discussions. The molecular dynamics simulations were performed on the Matterhorn Beowulf cluster at the Informatikdienste of the University of Zurich. We thank C. Bollinger and Dr A. Godknecht for their help in setting up and maintaining the cluster, and the Canton of Zurich for generous hardware support. This work was supported by the Swiss National Science Foundation.

Supplementary data

Supplementary data associated with this article can be found on doi:10.1016/j.jmb.2004.06.063.

References

1. Daggett, V. & Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* **28**, 18–25.
2. Bryngelson, J. & Wolynes, P. (1989). Intermediates and barrier crossing in a random energy-model (with applications to protein folding). *J. Phys. Chem.* **93**, 6902–6915.
3. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl Acad. Sci. USA*, **89**, 8721–8725.
4. Karplus, M. (1997). The Levinthal paradox: yesterday and today. *Fold. Des.* **2**, S69–S75.
5. Becker, O. M. & Karplus, M. (1997). The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **106**, 1495–1517.
6. Wales, D., Doye, J., Miller, M., Mortenson, P. & Walsh, T. (2000). Energy landscapes: from clusters to biomolecules. *Advan. Chem. Phys.* **115**, 1–111.
7. Krivov, S. V. & Karplus, M. (2002). Free energy disconnectivity graphs: application to peptide models. *J. Chem. Phys.* **117**, 10894–10903.
8. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
9. Albert, R., Jeong, H. & Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature*, **401**, 130–131.
10. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. ser. E*, **65**, 061910.1–061910.4.
11. Greene, L. H. & Higman, V. A. (2003). Uncovering network systems within protein structures. *J. Mol. Biol.* **334**, 781–791.
12. Newman, M. (2003). The structure and function of complex networks. *SIAM REV.* **45**, 167–256.
13. Scala, A., Amaral, L. A. N. & Barthelemy, M. (2001). Small-world networks and the conformation space of a short lattice polymer chain. *Europhys. Letters*, **55**, 594–600.
14. Doye, J. (2002). Network topology of a potential energy landscape: a static scale-free network. *Phys. Rev. Letters*, **88**, 238701.
15. De Alba, E., Santoro, J., Rico, M. & Jiménez, M. A. (1999). De novo, design of a monomeric three-stranded antiparallel β-sheet. *Protein Sci.* **8**, 854–865.
16. Ferrara, P., Apostolakis, J. & Caflisch, A. (2002). Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Struct., Funct. Genet.* **46**, 24–33.
17. Ferrara, P. & Caflisch, A. (2000). Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl Acad. Sci. USA*, **97**, 10780–10785.
18. Cavalli, A., Ferrara, P. & Caflisch, A. (2002). Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Struct., Funct. Genet.* **47**, 305–314.

19. Cavalli, A., Haberthür, U., Paci, E. & Caflisch, A. (2003). Fast protein folding on downhill energy landscape. *Protein Sci.* **12**, 1801–1803.
20. Ravasz, E. & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Phys. Rev. ser. E*, **67**, 026112.
21. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
22. Gsponer, J., Haberthür, U. & Caflisch, A. (2003). The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion sup35. *Proc. Natl Acad. Sci. USA*, **100**, 5154–5159.
23. Hiltpold, A., Ferrara, P., Gsponer, J. & Caflisch, A. (2000). Free energy surface of the helical peptide $Y(MEARA)_6$. *J. Phys. Chem. ser. B*, **104**, 10080–10086.
24. Gsponer, J. & Caflisch, A. (2001). Role of native topology investigated by multiple unfolding simulations of four SH3 domains. *J. Mol. Biol.* **309**, 285–298.
25. Gsponer, J. & Caflisch, A. (2002). Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl Acad. Sci. USA*, **99**, 6719–6724.
26. Ferrara, P., Apostolakis, J. & Caflisch, A. (2000). Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. ser. B*, **104**, 5000–5010.
27. Eaton, W. A., Munoz, V., Hagen, S., G, S., Jas, L. J., Lapidus, E. R. & Henry, J. (2000). Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327–359.
28. Andersen, C. A. F., Palmer, A. G., Brunak, S. & Rost, B. (2002). Continuum secondary structure captures protein flexibility. *Structure*, **10**, 174–184.
29. Du, R., Pande, V., Grosberg, A., Tanaka, T. & Shakhnovich, E. (1998). On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334–350.
30. Onuchic, J., Socci, N., Luthey-Schulten, Z. & Wolynes, P. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* **1**, 441–450.
31. Schonbrun, J. & Dill, K. A. (2003). Fast protein folding kinetics. *Proc. Natl Acad. Sci. USA*, **100**, 12678–12682.
32. Griffiths-Jones, S. R. & Searle, M. S. (2000). Structure, folding, and energetics of cooperative interactions between the β-strands of a *de novo* designed three-stranded antiparallel β-sheet peptide. *J. Am. Chem. Soc.* **122**, 8350–8356.
33. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways revealed and mapped. *Nature Struct. Biol.* **10**, 658–662.
34. Demarest, S. J., Hua, Y. X. & Raleigh, D. P. (1999). Local interactions drive the formation of nonnative structure in the denatured state of human alpha-lactalbumin: a high resolution structural characterization of a peptide model in aqueous solution. *Biochemistry*, **38**, 7380–7387.
35. Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.
36. Newman, M. (2001). The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA*, **98**, 404–409.
37. Chan, H. S. & Dill, K. A. (1998). Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins: Struct., Funct. Genet.* **30**, 2–33.
38. Karplus, M. (2000). Aspects of protein reaction dynamics: deviations from simple behaivor. *J. Phys. Chem. ser. B*, **104**, 11–27.
39. Lazaridis, T. & Karplus, M. (1997). "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, **278**, 1928–1931.
40. Koradi, R., Billeter, M. & Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

Edited by J. Thornton

(Received 23 March 2004; received in revised form 10 June 2004; accepted 15 June 2004)

Chapter 3

Local modularity measure for network clusterizations (PRE (2005) 72, 056107)

Local modularity measure for network clusterizations

Stefanie Muff, Francesco Rao, and Amedeo Caflisch
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zuerich, Switzerland
(Received 10 March 2005; revised manuscript received 9 September 2005; published 7 November 2005)

Many complex networks have an underlying modular structure, i.e., structural subunits (communities or clusters) characterized by highly interconnected nodes. The modularity Q has been introduced as a measure to assess the quality of clusterizations. Q has a global view, while in many real-world networks clusters are linked mainly *locally* among each other (*local cluster connectivity*). Here we introduce a measure of localized modularity LQ, which reflects local cluster structure. Optimization of Q and LQ on the clusterization of two biological networks shows that the localized modularity identifies more cohesive clusters, yielding a complementary view of higher granularity.

DOI: 10.1103/PhysRevE.72.056107 PACS number(s): 89.75.Hc

Complex networks are a powerful tool for the analysis of a diverse range of systems, including technological [1,2], social [3,4], and biological networks [5,6]. Especially in biology, thanks to high-throughput experiments, there is a tremendous growth of available data that can be efficiently analyzed and summarized in terms of complex networks [7,8]. In many cases, networks have an inherent modular structure which can represent functional units called communities or clusters, e.g., web pages of a certain subject [9], social groups [3,10], or biological modules [11,12]. However, there is neither an obvious and commonly accepted definition of communities nor a straightforward way to find the underlying modules of a network. Recently, many clustering algorithms have been proposed [13–18]. For a clusterization with K communities, the *modularity* $Q=\Sigma_{i=1}^{K}[e_{ii}-(a_i)_{in}(a_i)_{out}]$ has been introduced as a measure to assess the quality of a clusterization [19], where $e_{ii}=L_i/L_{tot}$, the effective fraction of links inside community i, is compared to $(a_i)_{in}(a_i)_{out}=(L_i)_{in}(L_i)_{out}/L_{tot}^2$ which is the predicted fraction of edges that fall into community i if the links in a directed network are set between nodes without regard to the community structure. Q is high when the clusterization is good and it can reach a maximum value of 1. Modularity is used to compare the quality of different clusterizations, e.g., to find the best split of a dendogram [20] or to validate different clusterization methods and furthermore as a fitness function in optimization procedures, where Q_{max} should correspond to the objectively best clusterization of a network [11,14]. The modularity is a global measure because the comparison of L_i/L_{tot} with $(L_i)_{in}(L_i)_{out}/L_{tot}^2$ assumes that connections between all pairs of nodes are equally probable, which reflects connectivity among all clusters.

On the other hand, in many complex networks most clusters are connected to only a small fraction of the remaining clusters. In metabolic networks, for instance, major pathways occur as clusters that are sparsely linked among each other [11]. Furthermore, in the protein folding network [6] communities are energy basins and transitions, i.e., connections, are allowed only between adjacent basins [15]. We call this property *local cluster connectivity*. In this paper, we introduce a measure for the quality of network clusterizations. To take into account local cluster connectivity and to overcome global network dependency, the approach of modularity is modified into a *local* version. The contribution to modularity for each cluster i is calculated for the subnetwork consisting of cluster i and its neighbor clusters. This requires the determination of i's neighborhood or, more precisely, all the links L_{i_N} that are contained in this neighborhood. The sum of the contributions of all K clusters yields

$$LQ=\sum_{i=1}^{K}\left[\frac{L_i}{L_{i_N}}-\frac{(L_i)_{in}(L_i)_{out}}{(L_{i_N})^2}\right].$$

We call LQ *localized modularity*. It is – in contrast to Q – not bounded by 1, but can take any value. The more locally connected clusters a network has, the higher LQ is. On the other hand, in a network where all communities are linked among each other, Q and LQ coincide.

It is interesting to compare the behavior of Q and LQ on different network topologies and use them as fitness functions for the optimization of network clusterizations [11,14]. We start with an illustration of the differences between Q and LQ by discussing a simple example of a scalable local cluster connectivity network, which we call the *school network* [Fig. 1(a)]. It is a toy model of social interactions between pupils in a school with l levels and c classes per level. Levels have periodic boundary conditions to avoid spurious boundary effects (in the first and last levels). In a real school, all the students of a class know each other and, as a first approximation, a student would interact most with people of his or her age. In the school network model, students are the nodes of the network and a link between two pupils is made if they know each other. Each class contains s fully connected students. A link between two students of the same level but different classes is placed with a (high) probability $p\leqslant 1$ and connections between students that are one level above or below [+1, Fig. 1(a)] are made with smaller probability $r<p$. No social interaction is assumed between persons that are more than one level apart from each other, i.e., if one of the students is more than one year older than the other [+2 or more, Fig. 1(a)]. Interestingly, when only two levels and two classes per level are considered, the school network model is essentially the same as the well-known (globally connected) four communities test network used in [11,14]. Hence, the

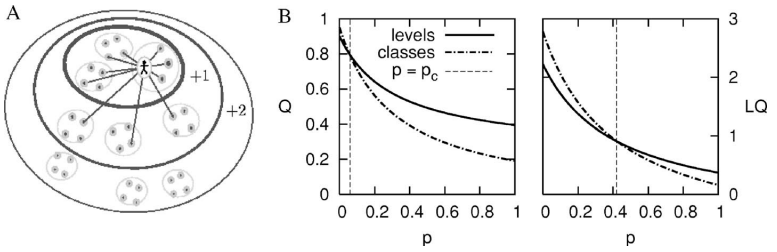

FIG. 1. (a) A student's view in the simplified schematic school network model with only three levels, three classes per level and four students per class: The student interacts with all his classmates, with other students on the same level with probability $p=0.5$, and with pupils one level above or below (+1) with probability $r=0.25$. No connections are assumed between students that are more than one level apart (+2 or more). (b) The p-dependent behavior of the modularity and the localized modularity in the school network with ten levels, two classes per level, 20 pupils per class, and $r=p/2$. The modularity favors the grouping of classes (solid line) in the same level for almost all p, whereas localized modularity favors communities consisting of single classes (dot-dashed line) for $p<0.42$.

school network is a simple generalization to locally connected networks. It is unweighted and undirected but an extension to directed and weighted networks, e.g., asymmetrical friendship, is straightforward.

A grouping of all the pupils on one level into the same cluster is reasonable for high p, i.e., when students of the same age interact among each other with high probability. But, as p decreases, classes become more and more separated from each other until they fully break apart for $p=0$, where a fitness measure is expected to favor clusterizations that identify classes. Therefore, we calculated modularity and localized modularity for the clusterization of nodes according to classes and according to levels for $p \in [0,1]$, $r=p/2$, and $s=20$ students per class. Figure 1(b) shows the Q and LQ values for ten levels and two classes per level. They were obtained analytically, using the expected numbers of links for each p. Both Q and LQ favor the clusterization into levels for p close to 1. LQ yields the same value for both clusterizations (*crossing point*) at $p_c^{LQ}=0.42$ and prefers the clusterization into classes for $p<0.42$. The modularity, on the other hand, has its crossing point at $p_c^{Q}=0.09$, i.e., it favors the classes only for $p<0.09$. In other words, Q considers the classes and not the levels as the best cluster partition only if the probability of interaction between two students of the same age but different classes is smaller than 10%.

The crossing point p_c depends on the number of levels and classes. Figure 2 shows the change of p_c upon variation of these two parameters with two, five, and ten classes per level, respectively (from top to bottom). It can be seen that p_c^{LQ} is higher than p_c^{Q} for all values of levels and classes, and is by construction constant for a fixed number of classes per level. On the other hand, p_c^{Q} strongly depends on network size which means that it favors different clusterizations as the number of levels increases, i.e., the lens of cluster detection becomes more coarse. Furthermore, it converges to 0 as l grows, meaning that Q favors the clusterization into levels for any $p \in [0,1]$, even though the classes on the same level are almost disconnected for small p.

These observations indicate that LQ is more reliable than Q to validate clusterizations in local cluster connectivity networks. The discrepancies between the two measures originate from the fact that Q compares the effective to the expected fraction of links in the clusters, no matter if a link is *possible* or not. The expected fraction of links is therefore underestimated in local cluster connectivity networks, thus the difference between the expected and the effective fraction of links (i.e., Q) is overestimated. On the other hand, LQ only takes into account local link expectations. Furthermore, note that modularity as high as 0.8 has been found in Erdös-Réni (ER) random graphs, scale-free networks, and regular lattices [21,22].

In recent years, biological networks [23] have attracted the attention of many scientists for their potential impact on the understanding of living systems. Metabolic and protein-protein interaction networks have been clustered by Q optimization [11] and the MCL method [24], respectively. To investigate the behavior of Q and LQ on real-world networks we optimized the clusterizations of two recent realizations of the metabolic and protein-protein interaction networks of E. coli by simulated annealing (SA), using each of the two measures as cost function. For each temperature T, $c_1 n^2$ single-node and $c_2 n$ multinode moves, like splitting and merging of (adjacent) communities, were performed, where $c_{1,2}$ are constants and n is the number of nodes in the network. Furthermore, T was iteratively reduced to $c_3 T$ with a constant

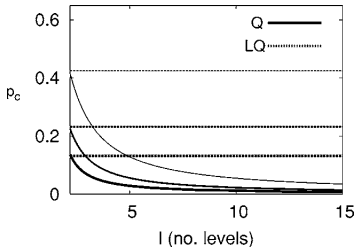

FIG. 2. Dependence of p_c on network size: for two, five, and ten classes per level (from top to bottom), p_c^{LQ} (dotted lines) is always higher than p_c^{Q} (solid lines) showing that LQ favors the clusterization into classes for higher p while Q almost always prefers the grouping into levels. Moreover, p_c^{Q} is rather sensitive on the size of the network and converges to 0 as the network grows, while p_c^{LQ} does not depend on the number of levels.

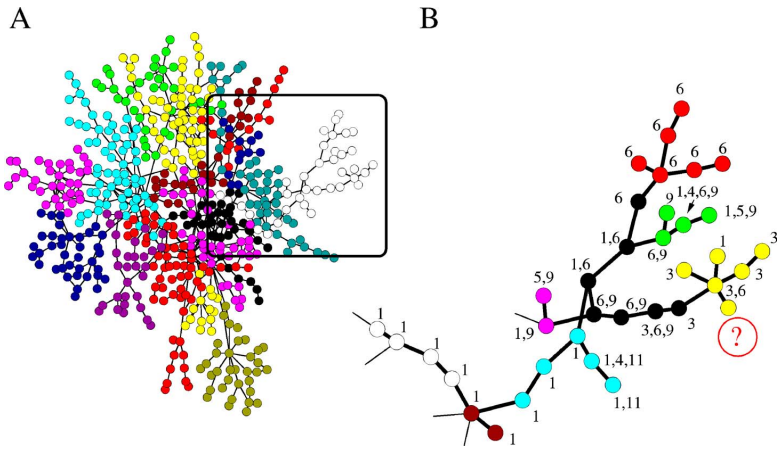

FIG. 3. (Color online) (a) Largest connected component of the metabolic network of *E. coli*. The coloring scheme represents the clusterization found by optimizing modularity. Some colors are used twice. (b) *LQ* clusterization of the white *Q* cluster with the annotation of different pathways. According to *LQ* it is highly probable that the unassigned yellow node (*N*-acetyl-alpha-*D*-glucosamine 1-phosphate, marked as "?") belongs to the carbohydrate metabolism (label 3).

$c_3 < 1$. This move set and cooling scheme is similar to the one used in [11]. The computational effort for the two measures scales as $O(K)$, even though the calculation of *LQ* is slightly more expensive since it involves the determination of neighborhoods for each cluster.

(i) *The metabolic network of E. coli*. We use the metabolic pathway database developed by Ma and Zeng [25], which has been derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [26]. Figure 3 shows the largest connected component of the *E. coli* metabolic network in this database. It contains 563 nodes and 708 links which have been treated undirected. Each node is assigned to between zero and nine out of 11 possible pathways. The optimization with fitness function *Q* leads to a division into 16 clusters consisting of 35 metabolites on average (as colored in Fig. 3) and takes a value as high as $Q_{max}=0.82$. On the other hand, *LQ* optimization leads to a maximum of $LQ_{max}=12.1$ with 132 clusters, each containing an average of 4.3 metabolites. The optimization of the two measures finds clusters at a different level, which yields complementary information. As expected, *Q* is based on a global view and depends on the size of the network. As a consequence, optimizing a network with more metabolites would lead to larger *Q* clusters. This problem is likely to arise because, as more data become available, the network and its largest connected component will grow. On the other hand, *LQ* finds the lowest-level modules, independent on the rest of the network. Still, a major motivation to find clusters is to obtain information about presumed pathways of nonannotated metabolites. Figure 3(b) zooms into one of the *Q* clusters (white) and shows the splitting into smaller *LQ* clusters. The numbers indicate the respective pathway(s) of the nodes. Note that an *LQ* cluster is not necessarily fully contained in a *Q* cluster, i.e., a smaller (local) cluster may be only *partially* contained in a larger one. In the considered cluster of Fig. 3(b), the further division is justified because it results in more homogeneous subclusters. The yellow community, for instance, contains mainly nodes belonging to the carbohydrate metabolism pathway (label 3). According to this, the unassigned node [*N*-acetyl-alpha-*D*-glucosamine 1-phosphate, labeled as "?" in Fig. 3(b)] can also be classified in pathway 3 with high confidence. This would have been impossible when considering the white cluster obtained by *Q* whose nodes are assigned mainly to pathway 6 (glycan biosynthesis and metabolism) and 1 (amino-acid metabolism).

To obtain a more quantitative analysis, we compute the conditioned probability

$$P[i,j] = P[\pi(i) \cap \pi(j) \neq \varnothing \,|\, c(i) = c(j)] \qquad (1)$$

that two nodes i and j, lying in the same cluster c, share at least one pathway (π). For the *Q* clusterization, this probability is $P_Q[i,j]=0.57$, while $P_{LQ}[i,j]=0.73$, reflecting the higher homogeneity of the *LQ* clusters. Comparison to the null case, where nodes are picked at random from the network, yields $P_R[i,j]=0.26$ and the probability that any pair of linked nodes shares a pathway is 0.59, thus essentially the same as for the clustering with *Q*.

(ii) *The protein-protein interaction (PPI) network of E. coli*. A set of 716 verified interactions involving 270 proteins of *E. coli* has been reported [27]. We again focused on the largest connected component consisting of 230 proteins and 695 undirected connections (Fig. 4). Identifying clusters can help to find indications about the function of unknown proteins. Again, modularity and localized modularity differ in the granularity of the clusters, similar to using two different lenses of a microscope. While the highest value for *Q* has been found for a clusterization with seven communities ($Q_{max}=0.49$), *LQ* splits the network into 56 communities ($LQ_{max}=2.97$). An example where *LQ* yields a more accurate "guess" is given in Fig. 4(b), where the *LQ* clusterization further subdivides the black cluster of Fig. 4(a). The proteins in the green circle are part of the DNA polymerase complex (dnaE, dnaQ, dnaX, dnaQ, holA, holB, holC, holD and holE). According to *LQ*, the unknown protein b1808 appears to be a protein of this complex. On the other hand, the black cluster obtained by *Q* is more heterogeneous which makes a functional assignment of b1808 difficult.

In conclusion, a measure for the quality of network clusterizations, called *localized modularity*, has been introduced and compared to the widely used *modularity*. Both measures can be used essentially in the same way. The latter has been

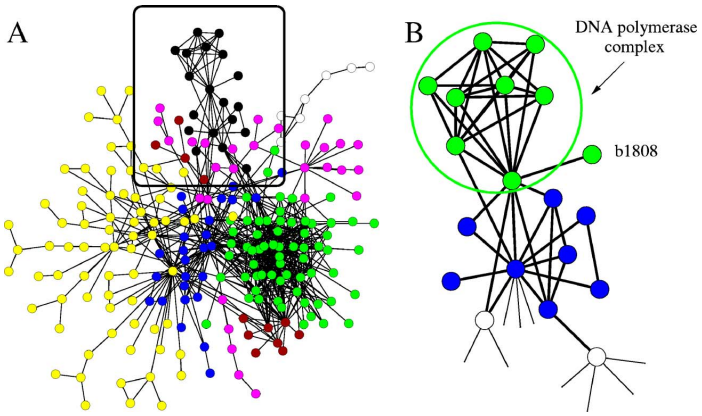

FIG. 4. (Color online) (a) Largest connected component of the PPI of *E. coli*. The colors represent the clusterization found by optimizing modularity. (b) *LQ* clusterization of the black *Q* cluster. The green circle contains proteins belonging to the DNA polymerase complex. The unknown protein b1808 is assigned to this complex according to *LQ* while the complete *Q* cluster is heterogeneous.

applied previously by others to assess the clusterization quality in many networks and has been used to find the best split of a dendogram and as fitness function in optimization algorithms. Finding clusters by optimizing a given fitness function has the advantage of not using any parameters (unlike many other clustering methods [15,17,18]). *Q* depends on global properties like the network size and the cluster connectivity. However, in many real-world networks, communities are merely connected locally, i.e., most pairs of clusters are not linked. We have called such organization *local cluster connectivity*. By detailed investigation of model networks as well as the optimization of *Q* and *LQ* on two biological networks, we have provided evidence that the two measures give a view of different depth into the cluster structure. In contrast to *Q*, *LQ* takes into account individual clusters and their nearest neighbors, generating high-confident clusters, irrespective of the rest of the network. Thus, the two measures provide complementary information. Furthermore, the *LQ* approach can be generalized to second or higher nearest neighbors which, albeit computationally more expensive, might yield additional insights, as if one were to use different lenses of a microscope.

This work was supported by a grant from the Swiss National Science Foundation.

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (2004).
[2] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[3] J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, London, 2000).
[4] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).
[5] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **427**, 839 (2004).
[6] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[7] Y. Xia, H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, H. Zhao, and M. Gerstein, Annu. Rev. Biochem. **73**, 1051 (2004).
[8] A.-L. Barabási and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[9] J.-P. Eckmann and E. Moses, Proc. Natl. Acad. Sci. U.S.A. **99**, 5825 (2002).
[10] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, UK, 1994).
[11] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[12] M. B. Eisen, P. T. Spellman, P. O. Brow, and D. Botstein, Proc. Natl. Acad. Sci. U.S.A. **95**, 14863 (1998).
[13] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[14] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[15] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[17] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, Nucleic Acids Res. **30**, 1575 (2002).
[18] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature (London) **435**, 814 (2004).
[19] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[20] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[21] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).
[22] C. P. Massen and J. P. K. Doye, Phys. Rev. E **71**, 046101 (2005).
[23] M. G. Grigorov, Drug Discovery Today **10**, 365 (2005).
[24] J. P. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, Proteins: Struct., Funct., Bioinf. **54**, 49 (2004).
[25] H. Ma and A.-P. Zeng, Bioinformatics **19**, 270 (2003).
[26] M. Kanehisa and S. Goto, Nucleic Acids Res. **28**, 27 (2000).
[27] G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, Nature (London) **433**, 531 (2005).

Chapter 4

Estimation of protein folding probability from equilibrium simulations (JCP (2005), 122, 184901)

THE JOURNAL OF CHEMICAL PHYSICS **122**, 184901 (2005)

Estimation of protein folding probability from equilibrium simulations

Francesco Rao, Giovanni Settanni, Enrico Guarnera, and Amedeo Caflisch[a)]
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 12 January 2005; accepted 23 February 2005; published online 6 May 2005)

The assumption that similar structures have similar folding probabilities (p_{fold}) leads naturally to a procedure to evaluate p_{fold} for every snapshot saved along an equilibrium folding-unfolding trajectory of a structured peptide or protein. The procedure utilizes a structurally homogeneous clustering and does not require any additional simulation. It can be used to detect multiple folding pathways as shown for a three-stranded antiparallel β-sheet peptide investigated by implicit solvent molecular dynamics simulations. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1893753]

I. INTRODUCTION

The folding probability p_{fold} of a protein conformation saved along a Monte Carlo or molecular dynamics (MD) trajectory is the probability to fold before unfolding.[1] It is a useful measure of kinetic distance from the folded, i.e., functional state, and can be used to validate transition state ensemble (TSE) structures, which should have $p_{\mathrm{fold}} \approx 0.5$. Such validation consists of starting a large number of trajectories from putative TSE structures with varying initial distribution of velocities and counting the number of those that fold within a "commitment" time which has to be chosen much longer than the shortest time scales of conformational fluctuations and much shorter than the average folding time.[2] The concept of p_{fold} calculation originates from a method for determining transmission coefficients, starting from a known transition state[3] and the identification of simpler transition states in protein dynamics (e.g., tyrosine ring flips).[4] The approach has been used to identify the otherwise very elusive folding TSE by atomistic Monte Carlo off-lattice simulations of small proteins with a *Go* potential,[2,5] as well as implicit solvent MD (Refs. 6 and 7) and Monte Carlo[8] simulations with a physicochemical based potential. The number of trial simulations needed for the reliable evaluation of p_{fold} makes the estimation of the folding probability computationally very expensive. For this reason, here we propose a method to estimate folding probabilities for *all* structures visited in an equilibrium folding-unfolding trajectory without any additional simulation.

II. METHODS

A. Molecular dynamics simulations

Beta3s is a designed 20-residue sequence whose solution conformation has been investigated by NMR spectroscopy.[9] The NMR data indicate that beta3s in aqueous solution forms a monomeric (up to more than 1 mM concentration) triple-stranded antiparallel β sheet, in equilibrium with the denatured state.[9] We have previously shown that in implicit solvent[10] molecular dynamics simulations beta3s folds reversibly to the NMR solution conformation, irrespective of the starting structure.[11] Recently, four molecular dynamics simulations of beta3s were performed at 330 K for a total simulation time of 12.6 μs.[12] There are 72 folding events and 73 unfolding events and the average time required to go from the denatured state to the folded conformation is 83 ns. The 12.6 μs of simulation length is about two orders of magnitude longer than the average folding or unfolding time, which are similar because at 330 K the native and denatured states are almost equally populated.[12] For the p_{fold} analysis the first 0.65 μs of each of the four simulations were neglected so that along the 10 μs of simulations there are a total of 500 000 snapshots because coordinates were saved every 20 ps.

The simulations were performed with the program CHARMM.[13] Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[13]). A mean field approximation based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute.[10] The two surface-tension-like parameters of the solvation model were optimized without using beta3s. The same force field and implicit solvent model have been used recently in molecular dynamics simulations of the early steps of ordered aggregation,[14] and folding of structured peptides,[10,11] as well as small proteins of about 60 residues.[15] Despite the absence of collisions with water molecules, in the simulations with implicit solvent the separation of time scales is comparable with that observed experimentally. Helices fold in about 1 ns,[16] β hairpins in about 10 ns,[16] and triple-stranded β sheets in about 100 ns,[12] while the experimental values are $\sim$0.1 μs,[17] $\sim$1 μs,[17] and $\sim$10 μs,[9] respectively.

B. Clusterization

The 500 000 conformations obtained from the simulations of beta3s (see above) were clustered by the leader algorithm.[18] Briefly, the first structure defines the first cluster and each subsequent structure is compared with the set of clusters found so far until the first similar structure is found. If the structural deviation (see below) from the first conformation of all of the known clusters exceeds a given thresh-

[a)]Author to whom correspondence should be addressed. FAX: +41 44 635 68 62. Electronic mail: caflisch@bioc.unizh.ch

old, a new cluster is defined. The leader algorithm is very fast even when analyzing large sets of structures such as in the present work. The results presented here were obtained with a structural comparison based on the distance root mean square (DRMS) deviation considering all distances involving C_α and/or C_β atoms and a cutoff of 1.2 Å. This yielded 78 183 clusters. The DRMS and root mean square deviation of atomic coordinates (upon optimal superposition) have been shown to be highly correlated.[2] The DRMS cutoff of 1.2 Å was chosen on the basis of the distribution of the pairwise DRMS values in a subsample of the wild-type trajectories. The distribution shows two main peaks that originate from intracluster and intercluster distances, respectively (data not shown). The cutoff is located at the minimum between the two peaks. The main findings of this work are valid also for clusterization based on secondary structure similarity.[7,19]

C. Folding probability

For the computation of p_{fold} a criterion (Φ) is needed to determine when the system reaches the folded state. Given a clusterization of the structures, a natural choice for Φ is the visit of the most populated cluster which for structured peptides and proteins is not degenerate (other criteria are also possible, e.g., fraction of native contacts Q larger than a given threshold). Given Φ and a commitment time (τ_{commit}), the folding probability $p_{\text{fold}}(i)$ of a MD snapshot i is computed as[1,2]

$$p_{\text{fold}}(i) = \frac{n_f(i)}{n_t(i)}, \tag{1}$$

where $n_f(i)$ and $n_t(i)$ are the number of trials started from snapshot i which reach within a time τ_{commit} the folded state and the total number of trials, respectively.

Every simulation started from snapshot i can be considered as a Bernoulli trial of a random variable θ with value 1 (folding within τ_{commit}) or 0 (no folding within τ_{commit}). The variable θ has average and variance on the average of the form

$$\langle\theta\rangle = p_{\text{fold}} = \frac{1}{n_t}\sum_{i=1}^{n_t}\theta_i,$$
$$\sigma^2_{\langle\theta\rangle} = \frac{1}{n_t}p_{\text{fold}}(1-p_{\text{fold}}), \tag{2}$$

where n_t is the total number of trials and the accuracy on the p_{fold} value increases with n_t.

In Fig. 1 the distribution of the first passage time (fpt) to the folded state is shown. The double peak shape of the distribution provides evidence for the different time scales between *intra*basin and *inter*basin transitions. A value of 5 ns is chosen for τ_{commit} because events with smaller time scales correspond to the diffusion within the native free-energy basin, while events with larger time scales are transitions from other basins to the native one, i.e., folding/unfolding events.[12]

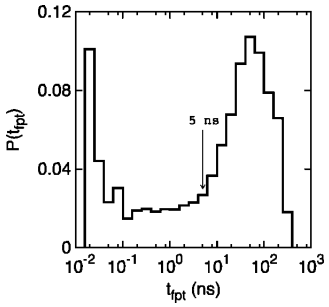

FIG. 1. Probability distribution for the first passage time (fpt) to the most populated cluster (*folded state*) of the DRMS 1.2 Å clusterization.

III. FOLDING PROBABILITY FROM EQUILIBRIUM TRAJECTORIES

The basic assumption of the present work is that conformations that are structurally similar have the same kinetic behavior, hence they have similar values of p_{fold}. Note that the opposite is not necessarily true as explained in Sec. IV for the TSE and the denatured state. To exploit this assumption, snapshots saved along a trajectory are grouped in structurally similar clusters.[20] Then the τ_{commit} segment of MD trajectory following each snapshot is analyzed to check if the folding condition Φ is met (i.e., the snapshot "folds"). For each cluster, the ratio between the snapshots which lead to folding and the total number of snapshots in the cluster is defined as the cluster $-p_{\text{fold}}$ (P_f^C; throughout the text uppercase P and lowercase p refer to folding probability for clusters and individual snapshots, respectively). This value is an approximation of the p_{fold} of any single structure in the cluster which is valid if the cluster consists of structurally similar conformations. In other words, the occurrence of the folding event for the snapshots of a given cluster can be considered as a Bernoulli trial of a random variable θ. The average of θ and variance on the average for the set of snapshots belonging to a given cluster α can be written as

$$P_f^C[\alpha] = \langle\theta\rangle = \frac{1}{W}\sum_{i=1}^{W}\theta_i, \quad i \in \alpha,$$
$$\sigma^2_{\langle\theta\rangle} = \frac{1}{W}P_f^C(1-P_f^C), \tag{3}$$

where W is the number of snapshots in cluster α. P_f^C is the average folding probability over a set of structurally homogeneous conformations. Using the clustering and the folding criterion Φ introduced above, values of P_f^C for the 78 183 clusters can be computed by Eq. (3), i.e., the number of conformations of the cluster that fold within 5 ns divided by the total number of conformations belonging to the cluster.

In this paper we provide evidence that the basic assumption mentioned above, that is, similar conformations have similar folding probabilities, holds in the case of beta3s, a three-stranded antiparallel β-sheet peptide investigated by MD.[12] Moreover, we show that the computationally expensive

TABLE I. DRMS clusters used for the calculation of P_f.

Cluster	P_f^C [a]	P_f [b]	$\sigma_{p_{\text{fold}}}$ [c]	N [d]	W [e]	W_{sample} [f]
1	0.00	0.03	0.04	150	144	15
2	0.11	0.05	0.06	150	449	15
3	0.06	0.05	0.07	120	36	12
4	0.08	0.07	0.08	140	555	14
5	0.10	0.08	0.06	100	10	10
6	0.13	0.12	0.18	160	911	16
7	0.25	0.16	0.07	80	4	4
8	0.23	0.20	0.31	150	141	15
9	0.21	0.22	0.15	140	178	14
10	0.12	0.23	0.20	120	48	12
11	0.57	0.25	0.14	140	14	14
12	0.05	0.27	0.19	100	19	10
13	0.23	0.29	0.38	140	391	14
14	0.08	0.30	0.15	120	12	12
15	0.72	0.35	0.23	130	129	13
16	0.19	0.38	0.18	130	26	13
17	0.38	0.44	0.39	160	16	16
18	0.38	0.51	0.28	160	16	16
19	0.65	0.60	0.29	100	20	10
20	0.57	0.61	0.35	70	7	7
21	0.48	0.63	0.32	140	27	14
22	0.74	0.65	0.40	140	539	14
23	0.68	0.66	0.18	140	28	14
24	0.38	0.71	0.24	130	13	13
25	0.50	0.72	0.20	100	2	2
26	0.82	0.76	0.31	170	17	17
27	0.50	0.78	0.14	120	12	12
28	0.78	0.78	0.22	180	18	18
29	0.70	0.79	0.19	130	189	13
30	0.77	0.79	0.17	150	30	15
31	0.85	0.81	0.11	130	13	13
32	0.91	0.83	0.20	140	401	14
33	0.90	0.85	0.27	100	20	10
34	0.85	0.85	0.10	120	48	12
35	0.94	0.88	0.13	170	1990	17
36	0.71	0.94	0.07	70	7	7
37	0.95	0.95	0.06	150	855	15

[a]Cluster $-p_{\text{fold}}$ [P_f^C, Eq. (3)].
[b]Traditional, i. e., computationally expensive P_f value [Eq. (4)].
[c]Standard deviation of p_{fold} in a cluster [Eq. (5)].
[d]Total number of trials used to evaluate P_f. For every structure $n_t=10$ trials were performed ($N=n_t\,W_{\text{sample}}$) except for clusters 7 and 25 for which 20 and 50 trials were performed, respectively.
[e]Number of snapshots in the cluster.
[f]Number of snapshots used to evaluate P_f. The W_{sample} subset was obtained by selecting structures in a cluster every $\lfloor W/W_{\text{sample}} \rfloor$ saved conformations.

$$P_f[\alpha] = \frac{1}{W}\sum_{i=1}^{W} p_{\text{fold}}(i), \quad i \in \alpha, \tag{4}$$

which is measured by starting several simulations from each snapshot i in the cluster α with W snapshots, is well approximated by P_f^C whose evaluation is straightforward.

To test the assumption that similar structures have similar p_{fold} and to compare the values of P_f^C with those obtained from the standard approach,[1] folding probabilities P_f were computed for the structures of 37 clusters by starting several 5 ns MD runs from each structure and counting those that fold [Eqs. (1) and (4)]. The 37 clusters chosen among the 78 183 include both high- and low-populated clusters with P_f^C values evenly distributed in the range between 0 and 1 (see Table I). In the case of large clusters a subset of snapshots is considered for the computation of P_f. In those cases W is replaced in Eq. (4) by $W_{\text{sample}} < W$ that is the number of snapshots involved in the calculation.

The standard deviation of p_{fold} in a cluster is computed as

$$\sigma_{p_{\text{fold}}} = \sqrt{\langle (p_{\text{fold}}(i) - P_f[\alpha])^2 \rangle_{i\in\alpha}}. \tag{5}$$

In the case of full kinetic inhomogeneity, i.e., random grouping of snapshots, the p_{fold} value for all snapshots in a given cluster will be equal to 0 or 1, indicating the coexistence (in the same cluster) of structures that either exclusively fold or unfold. In this case $\sigma_{p_{\text{fold}}}$ reflects the Bernoulli distribution.[19] Figure 2(a) shows that, even when only $n_t=10$ runs per snapshot are used to compute p_{fold}, $\sigma_{p_{\text{fold}}}$ values are not compatible with those of a Bernoulli distribution. Moreover the values of the standard deviation decrease when the number of trials n_t increases, as reported in Fig. 2(b) for two sample clusters. The asymptotic value of $\sigma_{p_{\text{fold}}}$ ($n_t \to \infty$) for these two data sets is of 0.05 and 0.2. This value cannot reach zero because snapshots in a cluster are similar but not identical. These results suggest that snapshots inside the same cluster are kinetically homogeneous and a statistical description of p_{fold} can be adopted, that is, folding probabilities are computed as cluster averages (instead of single snapshots) by means of P_f and P_f^C.

We still have to verify that P_f^C indeed approximates the computationally expensive P_f. Namely, for the 37 clusters mentioned above a correlation of 0.89 between P_f^C and P_f is found with a slope of 0.86 (see Fig. 3(a) and Table I), indicating that the procedure is able to estimate folding probabilities for clusters on the folding-transition barrier ($P_f \sim 0.5$) as well as in the folding ($P_f \sim 1.0$) or unfolding ($P_f \sim 0.0$) regions. The error bars for P_f^C in Fig. 3(a) are derived from the definition of variance given in Eq. (3). In the same spirit of Eq. (3) the folding probability P_f and its variance are written as

$$\begin{aligned} P_f &= \langle \theta \rangle = \frac{1}{N}\sum_{i=1}^{N} \theta_i, \\ \sigma^2_{\langle\theta\rangle} &= \frac{1}{N} P_f (1 - P_f), \end{aligned} \tag{6}$$

where $N=\Sigma n_t$ is the total number of runs and θ is equal to 1 or 0, if the run folded or unfolded, respectively. Note that the same number of runs n_t has been used for every snapshot of a cluster. The large vertical error bars in Fig. 3(a) correspond to clusters with less than ten snapshots. The largest deviations between P_f and P_f^C are around the 0.5 region. This is due to the limited number of crossings of the folding barrier observed in the MD simulation [Fig. 3(b), around 70 events of folding[12]]. Improvements in the accuracy for the estimation of P_f are achieved as the number of folding events, i.e., the simulation time, increases [Figs. 3(c)–3(e)].

The two main results of this study, i.e., the kinetic homogeneity of the clusters and the validity of P_f^C as an ap-

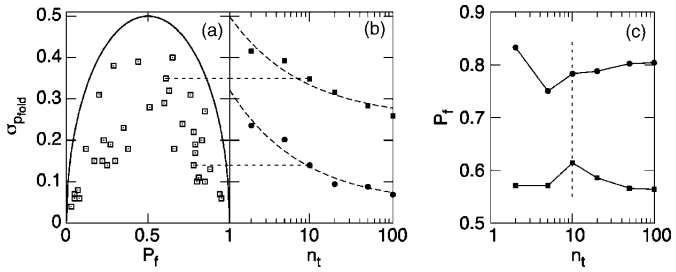

FIG. 2. Standard deviation $\sigma_{p_{\text{fold}}}=\sqrt{\langle(p_{\text{fold}}(i)-P_f[\alpha])^2\rangle_{i\in\alpha}}$ of the p_{fold} for the 37 DRMS clusters used in the study. (a) $\sigma_{p_{\text{fold}}}$ as a function of P_f compared to a Bernoulli distribution (solid line). Ten trials were performed for each snapshot. The largest values for the standard deviation are located around the 0.5 region and this is probably due to the Bernoulli process ($\theta=0,1$) used for the calculation of p_{fold}. (b) $\sigma_{p_{\text{fold}}}$ dependence on the number of trials used to evaluate p_{fold}. The dashed curves are fits with a $(a/\sqrt{x})+b$ function. The horizontal dashed lines are drawn to help identifying in (a) the two clusters used in (b). (c) Dependence of P_f on the number of trials n_t for the two clusters used in (b).

proximation of P_f, are robust with respect to the choice of the clusterization. Similar results can be obtained also with different flavors of conformation space partitioning, as long as they group together structurally homogeneous conformations, e.g., clusterization based on root mean square deviation of atomic coordinates (RMSD) or secondary structure strings.[19] The latter are appropriate for structured peptides but not for proteins with irregular secondary structure because of string degeneracy. Note that partitions based on order parameters (like native contacts) are usually unsatisfactory and not robust. This is mainly due to the fact that clusters defined in this way are characterized by large structural heterogeneities.[7]

IV. ANALYSIS OF TRANSITION STATE ENSEMBLE

The folding probability of structure i is estimated as $p_{\text{fold}}(i)=P_f^C[\alpha]$ for $i\in\alpha$. This approximation allows to plot the pairwise RMSD distribution of beta3s structures with $p_{\text{fold}}>0.51$ (native state), $0.49<p_{\text{fold}}<0.51$ (transition state ensemble, TSE), and $p_{\text{fold}}<0.49$ (denatured state) [Fig. 4(a)]. For the native state, the distribution is peaked around low values of RMSD ($\sim$1.5 Å) indicating that structures with $p_{\text{fold}}>0.51$ are structurally similar and belong to a nondegenerate state. The statistical weight of this group of structures is 49.4% and corresponds to the expected statistics for the native state because the simulations are performed at the melting temperature. In the case of TSE, the distribution is broad because of the coexistence of heterogeneous structures. This scenario is compatible with the presence of multiple folding pathways. Beta3s folding was already shown to involve two main average pathways depending on the sequence of formation of the two hairpins.[7,11] Here, a *naive* approach based on the number of native contacts[11] is used to structurally characterize the folding barrier. TSE structures

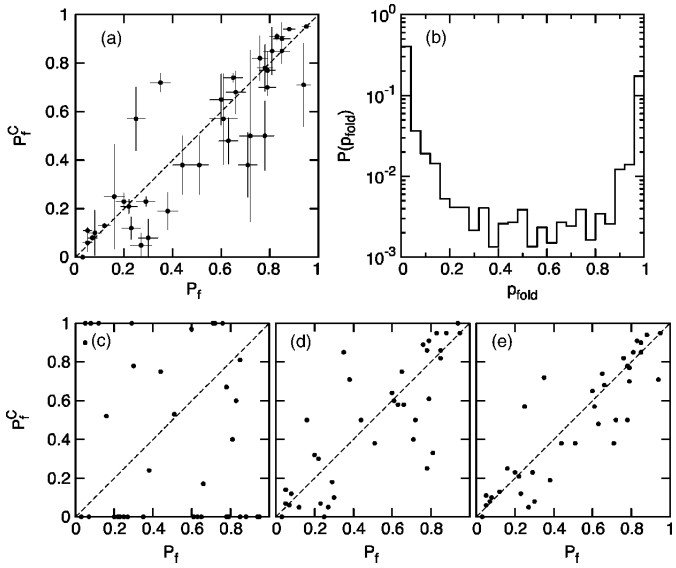

FIG. 3. Cluster folding probability P_f^C. (a) Scatter plot of P_f^C vs P_f. The DRMS 1.2 Å clusterization and the folding criterion Φ (reaching the most populated cluster within $\tau_{\text{commit}}=5$ ns) were used. (b) Probability distribution of the p_{fold} value for the 500 000 snapshots saved along the 10 μs MD trajectory. The folding probability for snapshot i is computed as $p_{\text{fold}}(i)=P_f^C[\alpha]$ for $i\in\alpha$. (c–e) Scatter plot of P_f^C vs P_f for 1.0, 5.0, and 10 μs of simulation time, respectively.

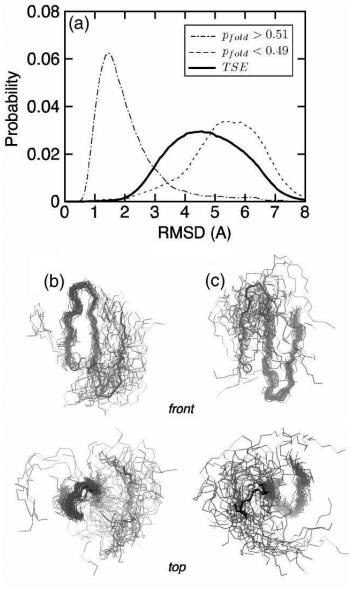

FIG. 4. Transition state ensemble (TSE) of beta3s. (a) RMSD pairwise distribution for structures with $p_{\mathrm{fold}} > 0.51$ (native state), $0.49 < p_{\mathrm{fold}} < 0.51$ (TSE), and $p_{\mathrm{fold}} < 0.49$ (denatured state). (b) Type I and (c) type II transition states (thin lines). Structures are superimposed on residues 2–11 and 10–19 with an average pairwise RMSD of 0.81 and 0.82 Å for type I and type II, respectively. For comparison, the native state is shown as a thick line with a circle to label the *N* terminus.

with number of native contacts of the first hairpin greater than the ones of the second hairpin are called type I conformations [Fig. 4(b)], otherwise they are called type II [Fig. 4(c)]. In both cases the transition state is characterized by the presence of one of the two native hairpins formed while the rest of the peptide is mainly unstructured. These findings are also in agreement with the complex network analysis of beta3s reported in Ref. 7. Finally, the denatured state shows a broad pairwise RMSD distribution around even larger values of RMSD (~5.5 Å), indicating the presence of highly heterogeneous conformations.

V. CONCLUSIONS

Two main results have emerged from the present study. First, snapshots grouped in structurally homogeneous clusters are characterized by similar values of p_{fold}. This result justifies the use of a statistical approach for the study of the kinetic properties of the structures sampled along a simulation. Second, given a set of structurally homogeneous clusters and a folding criterion, it is possible to obtain a first approximation of the folding probability for every structure sampled along an equilibrium folding-unfolding simulation. Thus, the cluster $-p_{\mathrm{fold}}$ is a quantitative measure of the kinetic distance from the native state and is computationally very cheap.[21] Furthermore, it can be used to detect multiple folding pathways. The accuracy in the identification of the transition state ensemble improves as the number of folding events observed in the simulation increases. Recently the cluster p_{fold} approach has been used to identify the transition state ensemble of a large set of beta3s mutants (for a total of 0.65 ms of simulation time[22]), which would have been impossible with traditional methods. As a further application, the cluster $-p_{\mathrm{fold}}$ procedure can be used to validate TSE conformations obtained by wide-spread $\overline{Go}$ models.

ACKNOWLEDGMENTS

The authors thank Stefanie Muff for useful and stimulating discussions and comments to the manuscript. They also thank Dr. Emanuele Paci for interesting discussions. The molecular dynamics simulations were performed on the Matterhorn Beowulf cluster at the Informatikdienste of the University of Zurich. They thank C. Bollinger, Dr. T. Steenbock, and Dr. A. Godknecht for setting up and maintaining the cluster. This work was supported by the Swiss National Science Foundation under Grant No. 205321-105946/1.

[1] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[2] I. Hubner, J. Shimada, and E. Shakhnovich, J. Mol. Biol. **336**, 745 (2004).
[3] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[4] S. H. Northrup, M. R. Pear, C. Y. Lee, J. A. McCammon, and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **79**, 4035 (1982).
[5] L. Li and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **98**, 13014 (2001).
[6] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **99**, 6719 (2002).
[7] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[8] P. Lenz, B. Zagrovic, J. Shapiro, and V. S. Pande, J. Chem. Phys. **120**, 6769 (2004).
[9] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez, Protein Sci. **8**, 854 (1999).
[10] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Genet. **46**, 24 (2002).
[11] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[12] A. Cavalli, U. Haberthür, E. Paci, and A. Caflisch, Protein Sci. **12**, 1801 (2003).
[13] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[14] J. Gsponer, U. Haberthür, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **100**, 5154 (2003).
[15] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).
[16] P. Ferrara, J. Apostolakis, and A. Caflisch, J. Phys. Chem. B **104**, 5000 (2000).
[17] W. A. Eaton, V. Munoz, J. Hagen, S G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, Annu. Rev. Biophys. Biomol. Struct. **29**, 327 (2000).
[18] J. A. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975).
[19] See EPAPS Document No. E-JCPSA6-122-505519 for Supplementary Material. This document can be reached via a direct link in the HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.html).
[20] Making a structural cluster analysis is equivalent to a partition of the conformation space. Given an appropriate partition, it is possible to analyze the dynamic behavior in terms of symbol sequences generated by the simulation. The symbol sequences describe the time evolution of the trajectories in a coarse-grained way. The mapping from the conformation space to the symbol space is called *symbolic dynamics* [C. Beck and F. Schloegl, *Thermodynamics of Chaotic Systems* (Cambridge University Press, Cambridge, 1993)]. Subsets α_i, also called *cells* or *clusters*, of different size and shape are used to partition the conformation space. The subsets are disjoint and cover the entire conformation space.
[21] The computation of P_f^C presented in this work takes few seconds on a desktop computer.
[22] G. Settanni, F. Rao, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **102**, 628 (2005).

Chapter 5

Φ−value analysis by molecular dynamics simulations of reversible folding (PNAS (2005), 102, 628)

Φ-Value analysis by molecular dynamics simulations of reversible folding

Giovanni Settanni*, Francesco Rao, and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Edited by Alan R. Fersht, University of Cambridge, Cambridge, United Kingdom, and approved December 2, 2004 (received for review September 11, 2004)

In Φ-value analysis, the effects of mutations on the folding kinetics are compared with the corresponding effects on thermodynamic stability to investigate the structure of the protein-folding transition state (TS). Here, molecular dynamics (MD) simulations (totaling 0.65 ms) have been performed for a large set of single-point mutants of a 20-residue three-stranded antiparallel β-sheet peptide. Between 57 and 120 folding events were sampled at near equilibrium for each mutant, allowing for accurate estimates of folding/unfolding rates and stability changes. The Φ values calculated from folding and unfolding rates extracted from the MD trajectories are reliable if the stability loss upon mutation is larger than ≈0.6 kcal/mol, which is observed for 8 of the 32 single-point mutants. The same heterogeneity of the TS of the wild type was found in the mutated peptides, showing two possible pathways for folding. Single-point mutations can induce significant TS shifts not always detected by Φ-value analysis. Specific nonnative interactions at the TS were observed in most of the peptides studied here. The interpretation of Φ values based on the ratio of atomic contacts at the TS over the native state, which has been used in the past in MD and Monte Carlo simulations, is in agreement with the TS structures of wild-type peptide. However, Φ values tend to overestimate the nativeness of the TS ensemble, when interpreted neglecting the nonnative interactions.

peptide folding | transition state

The Φ-value analysis is a protein engineering approach to investigate the transition state (TS) ensemble in protein folding (1, 2). The Φ value of residue i, that is the ratio $\Delta\Delta G_{TS\text{-}D}/\Delta\Delta G_{N\text{-}D}$ between the free energy change in the TS and native state (N) because of a mutation of the residue i [taking the denatured state (D) as a reference], represents the degree of nativeness of the structure around residue i in the TS. Observations derived from Φ-value analysis of many proteins, carried out in several research groups, have revealed that the TS is an ensemble of structures with an overall topology similar to the folded state, but with looser interactions (ref. 3 and references therein).

Φ values are usually interpreted in terms of native contacts (4). This description has been successfully used to obtain sets of conformations from the TS ensemble of several proteins (5–9) and to bias molecular dynamics (MD) trajectories toward the TS (10). On the other hand, specific nonnative interactions may be formed at both the TS and denatured-state ensemble and lead to a wrong picture of TS if not taken into account (11). Furthermore, different experimental conditions or mutations may determine detectable changes in the TS structure, showing the presence of parallel pathways (12, 13) and, thus, a heterogeneous TS. In addition, the ensemble average associated with the use of certain folding observables, like the degree of tryptophan burial, may disguise the presence of multiple folding pathways and folding intermediates (14). Namely, a recent study (15) suggests that not all conformations obtained in MD simulations by using Φ values as restraints on a subset of the native contacts belong to the TS.

The TS structures can be identified by MD simulations through the calculation of their folding probability P_{fold} (16), i.e., the probability that a trajectory started from a given structure reaches the folded state before unfolding. The concept of P_{fold} calculation was first introduced in a method for determining transmission coefficients, starting from a known TS (17), and used to identify TSs of simple conformational changes (e.g., tyrosine ring flips) (18). The approach has recently been used to study the otherwise very elusive folding TS by atomistic Monte Carlo off-lattice simulations of small proteins with a Go potential (6, 15) and a 21-residue polyalanine helix without Go potential (19) as well as by implicit solvent MD simulations with a physicochemical potential (8, 20). MD simulations are particularly useful to investigate structured peptides at atomic level of detail. Structured peptides usually form stable secondary structure elements, i.e., the building blocks of most of the larger proteins. Hence, they represent the simplest protein conformations. Understanding their process of folding will help to characterize the folding mechanism of larger proteins.

Here, we use MD simulations with an implicit model of the solvent to describe the TS ensemble and evaluate Φ values for several single-point mutants of Beta3s, a designed three-stranded antiparallel β-sheet peptide of 20 residues (21). Beta3s has been successfully characterized by MD simulations of reversible folding in which the native long-range nuclear Overhauser effect distance restraints are mostly satisfied (22). The length of the simulations in the present work has been chosen to achieve near-equilibrium sampling of the phase space of the peptides at the melting temperature of the wild type.

This work was inspired by the following questions: Is it possible to extract Φ values from trajectories near equilibrium? Are Φ values a measure of the extent of formation of contacts in the TS ensemble? How heterogeneous is the TS ensemble of a small structured peptide? Does the Φ-value analysis allow for the observation of any TS movement? What is the importance of nonnative contacts in the TS conformations? Analysis of the trajectories of Beta3s and its mutants allows for an atomic detailed picture of its phase space that is useful in answering these questions. In addition, the simulation results indicate that for the accuracy of a Φ value the threshold in the change of stability (0.6 kcal/mol) is smaller than postulated by Sanchez and Kiefhaber (1.7 kcal/mol) (23) and the same as suggested recently by Fersht and Sato (24).

Methods

Mutants of Beta3s. Thirty-two single-point mutations of the hydrophobic and aromatic side chains W2, I3, W10, Y11, I18, and Y19 were investigated (Fig. 1). The six sites of mutation are distributed along the sequence of the peptide, two for each strand. Between four and eight mutations have been studied for each site. Six of the 32 mutations are nondisruptive (I3A, I3V, Y11F, I18A, I18V, and Y19F), six mutations are conservative but change the steric properties of the side chain (I3M, Y11L, Y11M, I18M, Y19L, and Y19M), and the remaining 20 mutations are radical but acceptable because, in most of the cases, they do not change significantly the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: TS, transition state; MD, molecular dynamics; HB, hydrogen bond.

*To whom correspondence may be addressed. E-mail: caflisch@bioc.unizh.ch or settanni@bioc.unizh.ch.

www.pnas.org/cgi/doi/10.1073/pnas.0406754102

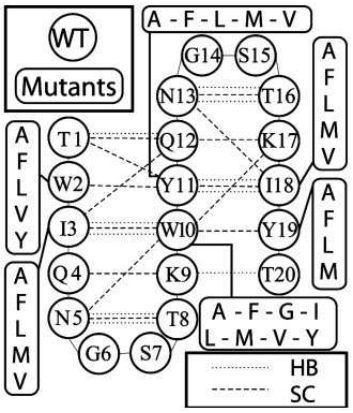

Fig. 1. Schematic representation of the Beta3s peptide, where the wild-type (WT) sequence and the mutants are indicated. The backbone HBs (dotted lines) and side-chain contacts (SC, dashed lines) common to most of the peptides are reported.

TS of the peptide, as showed in *Results and Discussion*. This result is probably due to the fact that the side chains of Beta3s are not fully buried in a densely packed hydrophobic core, as is the case in larger proteins (24).

MD Simulations. All simulations and part of the analysis of the trajectories were performed with the program CHARMM (25). Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms [PARAM19 force field (25)]. A mean field approximation based on the solvent-accessible surface was used to describe the main effects of the aqueous solvent on the solute (26). Ten MD runs of 2 μs each (total of 20 μs for each mutant) with different initial velocities were performed with the Berendsen thermostat at 330 K, which is close to the melting temperature of wild-type Beta3s (27). To improve sampling, the solute-solvent friction has been neglected that has no effect on the thermodynamic properties of the system (27). Despite the absence of collisions with water molecules, in the simulations with implicit solvent, relative rates are comparable with the values observed experimentally. Helices fold in $\approx$1 ns (28), β-hairpins in $\approx$10 ns (28) and triple-stranded β-sheets in $\approx$100 ns (27), whereas the experimental values are $\approx$0.1 (29), $\approx$1 (29), and $\approx$10 μs (21), respectively. Moreover, the effects of the viscosity on the folding and unfolding rates are essentially the same because the solvent-accessible surface and radius of gyration of Beta3s are only marginally larger in the 330 K denatured-state ensemble with respect to the native state (30). A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 10^6 conformations for each mutant. During the 20-μs simulation time, between 57 and 120 folding events were observed for every mutant (Table 1), thus providing sufficient statistical sampling for the kinetic analysis (see below for definition of folding event). This result is supported by the small difference in the native population measured for each individual mutant on two disjoint equal-size subsets of the trajectories (5% on average, the largest being 13%).

Clustering. The conformations of each peptide were clustered by the leader algorithm (31, 32) based on the distance rms (drms) deviation considering the C^α and C^β atoms. The drms and rms deviations were recently shown to be highly correlated (15). This algorithm is very fast, even when analyzing sets of 10^6 structures like in the present work. The drms cutoff of 1.2 Å has been chosen on the basis of the distribution of the pairwise drms values in a subsample of the wild-type trajectories. The distribution shows two main peaks that originate from intra- and intercluster distances, respectively (data not shown). The cutoff is located at the minimum between the two peaks.

Native Contacts. As in our previous work (22), a hydrogen bond (HB) is defined as native if the distance between the hydrogen and oxygen atoms is <2.5 Å for more than two-thirds of the conformations belonging to the most populated cluster. A side-chain contact is defined as native if the distance between the center of mass of the two residues averaged over the most populated cluster is <6.5 Å. Seventeen native contacts are common to the wild type and all mutants (but Y11V, see *Results and Discussion*) and 24 are common to the wild type and more than half of the mutants (Fig. 1). The latter set of contacts has been chosen as the reference for assessing the degree of nativeness of the structures, measured by the fraction of native contacts (Q). The high number of common native contacts shows that the most populated cluster of each mutant (except Y11V) is structurally the same as the one of the wild type.

Folding/Unfolding Events and Rates. The fraction of native contacts Q has been computed along the trajectories of all peptides. A folding (unfolding) event occurs when, along the trajectory, Q first reaches values >0.85 (<0.15) immediately after a previous unfolding (folding) event (22). All of the trajectories are started from the folded state, thus, the first event is always an unfolding. The average time separation between a folding (unfolding) event and the previous unfolding (folding) event is the folding (unfolding) time τ_f (τ_u). The folding and unfolding rates are $k_f = 1/\tau_f$ and $k_u = 1/\tau_u$, respectively.

Φ Values Calculated from Folding/Unfolding Rates. As in the kinetic experiments used to measure Φ^{exp} values, free energy changes with respect to wild type are computed from the folding and unfolding rates with the free energy of the denatured state as reference.

$$\Delta\Delta G^{kin}_{TS-D} = RT \log (k_f^{WT}/k_f^{mut}) \quad [1]$$

$$\Delta\Delta G^{kin}_{N-D} = RT \log [(k_f^{WT}/k_f^{mut})\cdot(k_u^{mut}/k_u^{WT})] \quad [2]$$

The Φ value is $\Phi = \Delta\Delta G^{kin}_{TS-D}/\Delta\Delta G^{kin}_{N-D}$. Values of $\Delta\Delta G^{kin}_{TS-D}$ and $\Delta\Delta G^{kin}_{N-D}$ from multiple mutations at the same site can be displayed on a single plot. The slope of the corresponding regression line is called the multipoint Φ value (23, 24).

Folding Probability and Definition of Native, TS, and Denatured-State Ensemble. The native state of the peptides consists of rapidly interconverting clusters, and the same holds for the denatured state. The following approach is used to group them together. The segment of MD trajectory after each snapshot is analyzed until it first reaches a Q value of >0.85 (i.e., the snapshot leads to folding) or <0.15 (unfolding). For each cluster, the ratio between the snapshots that lead to folding and the total number of snapshots in the cluster is defined as the cluster P_{fold}. This value is assumed as an approximation of the P_{fold} of any single structure of the cluster. We have recently shown that cluster P_{fold} values evaluated with this procedure correlate well with the P_{fold} values estimated by starting several MD simulations from different structures of a given cluster and counting the fraction of those that fold (F.R., G.S., and A.C., unpublished work and *Supporting Text,* which is published as supporting information on the PNAS web site).

The native state, the TS, and the denatured-state ensemble consist of the snapshots in the clusters with $P_{fold} \geq 0.51$, $0.49 \leq P_{fold} < 0.51$, and $P_{fold} < 0.49$, respectively (see Figs. 7 and 8, which are published as supporting information on the PNAS web site). Their statistical weights are W_N, W_{TS}, and W_D, respectively; these values

Table 1. Stability, folding/unfolding rates, and Φ values of the mutants

Mutation*	W_{highQ},† %	Nat. Cont.‡	W_{lowQ},§ %	τ_f,¶ ns	N_f‖	τ_u,** ns	N_u††	$\Delta\Delta G^{kin}_{N-D}$, kcal/mol‡‡	$\Delta\Delta G^{kin}_{TS-D}$, kcal/mol‡‡	Φ‡‡§§
WT	21.4	19.3 ± 1.7	2.9	70 ± 10	92	67 ± 6	94			
W2A	26.5	18.1 ± 2.3	3.5	107 ± 14	108	63 ± 6	114	−0.32 ± 0.15	−0.28 ± 0.13	0.87 ± 0.57
W2F	33.5	18.8 ± 2.2	3.4	106 ± 14	97	82 ± 8	103	−0.14 ± 0.16	−0.27 ± 0.13	—
W2L	24.9	18.2 ± 2.2	6.3	109 ± 16	101	63 ± 5	111	−0.34 ± 0.16	−0.30 ± 0.14	0.87 ± 0.57
W2V	23.6	18.3 ± 2.3	4.4	124 ± 17	95	62 ± 6	102	−0.43 ± 0.16	−0.38 ± 0.13	0.89 ± 0.45
W2Y	21.9	18.5 ± 2.4	6.4	129 ± 21	93	65 ± 6	98	−0.43 ± 0.16	−0.41 ± 0.14	0.95 ± 0.49
I3A	19.9	18.7 ± 2.2	3.9	137 ± 18	92	64 ± 5	101	−0.48 ± 0.15	−0.44 ± 0.13	0.93 ± 0.40
I3F	33.0	18.8 ± 2.1	3.3	121 ± 22	83	93 ± 8	91	−0.15 ± 0.17	−0.36 ± 0.15	—
I3L	28.5	18.5 ± 2.4	3.9	119 ± 19	94	72 ± 7	101	−0.31 ± 0.17	−0.35 ± 0.14	1.1 ± 0.77
I3M	30.2	18.9 ± 2.2	5.4	108 ± 19	94	81 ± 9	102	−0.16 ± 0.17	−0.29 ± 0.15	—
I3V	37.2	18.6 ± 2.1	5.2	124 ± 18	75	109 ± 10	83	−0.06 ± 0.16	−0.38 ± 0.14	—
W10A	31.8	19.5 ± 2.1	5.0	161 ± 21	74	95 ± 10	79	−0.32 ± 0.16	−0.55 ± 0.13	1.7 ± 0.93
W10F	41.3	18.7 ± 2.2	3.8	77 ± 9	120	78 ± 6	127	0.04 ± 0.14	−0.06 ± 0.12	—
W10G	12.7	19.3 ± 2.2	3.1	212 ± 32	60	68 ± 9	69	**−0.72** ± 0.17	−0.73 ± 0.14	**1.0 ± 0.31**
W10I	30.8	18.3 ± 2.1	6.0	129 ± 17	77	88 ± 9	83	−0.23 ± 0.16	−0.40 ± 0.13	—
W10L	20.8	18.8 ± 2.2	4.2	166 ± 22	81	58 ± 5	87	**−0.67** ± 0.16	−0.57 ± 0.13	**0.86 ± 0.28**
W10M	18.4	19.0 ± 2.2	6.6	155 ± 21	82	52 ± 5	91	**−0.68** ± 0.16	−0.52 ± 0.13	**0.76** ± 0.26
W10V	17.2	17.8 ± 2.5	6.7	259 ± 40	57	65 ± 11	64	**−0.88** ± 0.19	−0.86 ± 0.14	**0.98 ± 0.26**
W10Y	26.2	19.0 ± 2.1	3.5	118 ± 15	94	77 ± 7	98	−0.26 ± 0.15	−0.35 ± 0.13	—
Y11A	5.7	18.1 ± 2.0	2.3	249 ± 38	64	30 ± 3	71	**−1.37** ± 0.17	−0.84 ± 0.14	**0.61 ± 0.13**
Y11F	33.1	19.1 ± 2.2	4.4	138 ± 20	73	112 ± 12	79	−0.11 ± 0.16	−0.45 ± 0.14	—
Y11L	14.8	18.6 ± 2.1	4.8	169 ± 23	76	54 ± 6	83	**−0.72** ± 0.16	−0.58 ± 0.13	**0.81 ± 0.26**
Y11M	11.3	18.0 ± 2.2	3.5	152 ± 24	95	35 ± 3	105	**−0.94** ± 0.16	−0.51 ± 0.14	**0.54 ± 0.18**
Y11V	5.7	17.0 ± 2.7	7.4							
I18A	12.3	18.5 ± 2.3	2.4	168 ± 22	80	53 ± 6	88	**−0.73** ± 0.16	−0.58 ± 0.13	**0.79 ± 0.25**
I18F	21.3	19.0 ± 2.0	3.2	159 ± 23	74	72 ± 8	83	−0.50 ± 0.17	−0.54 ± 0.14	1.1 ± 0.46
I18L	22.2	19.0 ± 2.2	4.4	145 ± 19	73	94 ± 9	81	−0.26 ± 0.16	−0.48 ± 0.13	—
I18M	28.9	18.8 ± 2.2	4.8	97 ± 15	99	77 ± 6	106	−0.13 ± 0.16	−0.22 ± 0.14	—
I18V	29.6	18.8 ± 2.3	3.2	124 ± 20	87	86 ± 9	93	−0.22 ± 0.17	−0.38 ± 0.14	—
Y19A	20.7	18.6 ± 2.4	7.4	123 ± 18	90	84 ± 8	95	−0.23 ± 0.16	−0.37 ± 0.14	—
Y19F	29.2	18.4 ± 2.2	3.8	130 ± 18	92	71 ± 7	98	−0.37 ± 0.16	−0.41 ± 0.13	1.1 ± 0.59
Y19L	30.0	18.3 ± 2.2	3.2	117 ± 17	83	88 ± 8	89	−0.17 ± 0.16	−0.34 ± 0.13	—
Y19M	17.5	18.5 ± 2.3	6.2	155 ± 26	68	97 ± 10	76	−0.28 ± 0.17	−0.52 ± 0.15	—

*Mutants in italics are radical but acceptable and mutations in Roman are conservative (see *Methods* and ref. 24).
†Statistical weight of the three most populated clusters with $Q \geq 16/24$.
‡Average number of contacts in the three most populated clusters with $Q \geq 16/24$.
§Statistical weight of the three most populated clusters with $Q < 16/24$.
¶Average folding time.
‖Number of folding events.
**Average unfolding time.
††Number of unfolding events.
‡‡The SD have been obtained by propagation of the error on τ_f and τ_u.
§§Dashes indicate unreliable Φ values because of $|\Delta\Delta G^{kin}_{N-D}| < 0.3$ kcal/mol. The reliable Φ values and the corresponding large stability changes (24) are bold. The multipoint Φ values are **0.77**, **0.60**, **0.79**, **0.46**, **0.72**, and **1.23** for W2, I3, W10, Y11, I18, and Y19, respectively.

can be used to evaluate relative free energies by a different equation with respect to the kinetically evaluated $\Delta\Delta G^{kin}$. In the canonical ensemble, $\Delta G^{eq}_{TS-D} = -RT\log(W_{TS}/W_D)$ and $\Delta\Delta G^{eq}_{N-D} = -RT\log(W_N/W_D)$. An excellent match is observed between the $\Delta\Delta G^{kin}_{N-D}$ and $\Delta\Delta G^{eq}_{N-D}$ values (correlation coefficient of 0.99) and a good correlation between $\Delta\Delta G^{kin}_{TS-D}$ and $\Delta\Delta G^{eq}_{TS-D}$ (correlation coefficient of 0.73) (See Fig. 8). The agreement represents a consistency check for the parameters used to define folding and unfolding events. That activation free energy differences computed with the two sets of data show larger discrepancies than do changes in stability is because of the difficulty in sampling the TS ensemble. Note that the $\Delta\Delta G^{kin}_{TS-D}$ vs. $\Delta\Delta G^{eq}_{TS-D}$ correlation increases by decreasing until 0.02 the interval width of cluster P_{fold} values defining the TS ensemble (data not shown). The $\Delta\Delta G^{eq}_{TS-D}$ is only very slightly affected by the width of this interval because of the much larger number of structures in the denatured and native states than in the TS.

Structural Φ Values Based on Atomic Contacts. In each snapshot, a van der Waals contact is defined when the distance between two heavy atoms is <6 Å. $p_N(i)$ and $p_{TS}(i)$ measure the fraction of native and TS structures, respectively, in which the contact i is formed. If $p_N(i) > 0.66$, the contact i belongs to the set of the native contacts (NC). The structural Φ value

$$S_{Nat}\Phi(R) = \frac{1}{M_{NC(R)}} \frac{\sum_{i \in NC(R)} p_{TS}(i)}{\sum_{i \in NC(R)} p_N(i)}, \quad [3]$$

where $M_{NC(R)}$ is the number of native contacts of residue R, represents an estimate of the degree of nativeness of residue R at the TS ensemble. This measure has been used in the past to give a structural interpretation to experimental Φ values (4, 5, 10). An estimate of the relevance of nonnative interactions at the TS is obtained by extending the computation to all possible contacts (AC), including contacts not present in the NC set

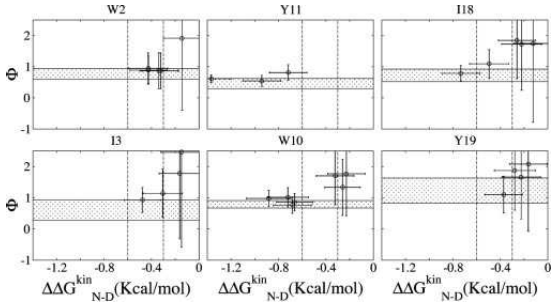

Fig. 2. Φ values as a function of change in the native state stability upon mutation. The shadowed horizontal region indicates 1 SD around the multipoint Φ value. The Φ values span a wide range and become anomalous for $|\Delta\Delta G^{kin}_{N-D}|$ smaller than ≈0.3 kcal/mol. The Φ values corresponding to mutations with $|\Delta\Delta G^{kin}_{N-D}| > 0.3$ are mainly in the normal range, i.e., between 0 and 1, and are in agreement with the multipoint Φ value. Vertical dashed lines are drawn at $\Delta\Delta G^{kin}_{N-D} = -0.3$ kcal/mol and $\Delta\Delta G^{kin}_{N-D} = -0.6$ kcal/mol. The Φ value of mutations I3V, W10F, and Y11F are located outside of the plot boundaries. The graphs are ordered according to the antiparallel β-sheet topology of Beta3s with vertical orientation of the three strands, and the N (*Left Upper*) and C (*Right Lower*) termini, respectively.

$$S_{\mathrm{All}}\Phi(R) = \frac{1}{M_{AC(R)}} \frac{\sum_{i \in AC(R)} p_{TS}(i)}{\sum_{i \in AC(R)} p_N(i)}, \qquad [4]$$

Results and Discussion

MD Simulations of Reversible Folding. The native structure of the wild type, i.e., the three-stranded antiparallel β-sheet with turns at G6-S7 and G14-S15, is also the most populated in all of the mutants, as shown by the cluster analysis of the trajectories (Table 1). The only exception is Y11V, which has a more distorted native state and has not been considered for further analysis. Moreover, there is no predominant structure in the denatured state for any of the mutants. The number of folding and unfolding events observed along the trajectories ranges from 57 to 120 and from 64 to 127, respectively (Table 1). Interestingly, the values of the stability change upon mutation, calculated with Eq. **2**, show that all mutants are less stable than wild-type Beta3s, except for W10F and I3V, which are essentially as stable as Beta3s. This result is not unexpected because Beta3s is a designed peptide whose sequence was carefully optimized for its fold (21).

Accuracy of Two-Point and Multipoint Φ Values. Fig. 2 shows the Φ values extracted from the simulations as a function of the change in free energy of folding upon mutation (see also Table 1). Because of the difficulties in the interpretation of Φ values, as many mutants as possible have been considered and the resulting Φ values divided into classes of reliable, tolerable, and unreliable, according to the size of the induced stability change $\Delta\Delta G^{kin}_{N-D}$. The deviations from the 0–1 range are large for unreliable Φ values, i.e., for mutations with $|\Delta\Delta G^{kin}_{N-D}| < 0.3$ kcal/mol, in agreement with previous observations (23). Indeed, in the unreliable class, the deviation can be observed for both radical mutations (e.g., I3F, W10A, and Y19A) and for nondisruptive mutations (e.g., I3V, Y11F, and I18V). For tolerable Φ values, i.e., $0.3 \text{ kcal/mol} \leq |\Delta\Delta G^{kin}_{N-D}| < 0.6$ kcal/mol, the deviation from the 0–1 interval is less frequent but the relative error is large. The eight reliable Φ values ($|\Delta\Delta G^{kin}_{N-D}| \geq 0.6$ kcal/mol) are all in the range of 0–1 and have a small SD. In a small structured peptide like Beta3s, most residues have a relatively large exposed surface area in the folded state so that conservative mutations generally induce small free-energy changes. Indeed, among the six conservative mutations, only I18A falls into the reliable class. For this reason, more radical mutations have been also investigated.

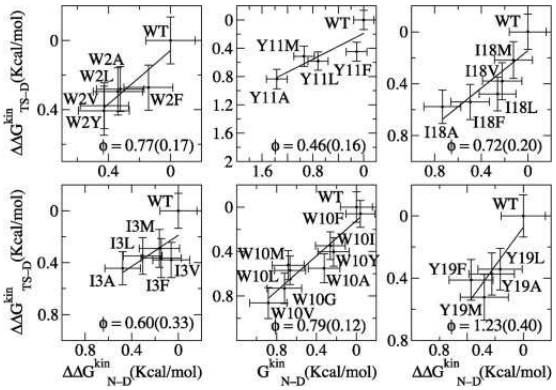

Fig. 3. $\Delta\Delta G^{kin}_{TS-D}$ plotted vs. $\Delta\Delta G^{kin}_{N-D}$ for all of the mutants grouped according to the mutation site along the structure of Beta3s. The optimal regression line (including the wild-type data point) is plotted, and its slope, i.e., the multipoint Φ value, is reported in the lower right corner of each graph with the SD derived from the fit in parentheses. The correlation coefficient is 0.91, 0.67, 0.93, 0.86, 0.87, and 0.88 for W2, I3, W10, Y11, I18, and Y19 mutants, respectively.

The multipoint Φ of Beta3s as extracted from the simulations are reported in Fig. 3. The good linear relationship between $\Delta\Delta G^{kin}_{TS-D}$ and $\Delta\Delta G^{kin}_{N-D}$, observed in mutants of W2, W10, Y11, and Y19, supports the validity of the multipoint analysis for these residues and indicates a substantial similarity among the folding TS ensembles of those peptides. In mutants of I3, the linear correlation is weaker than the others, and in I18, there is a change in the slope for $\Delta\Delta G^{kin}_{N-D} < -0.3$ kcal/mol. A possible explanation for the presence of a linear relationship in the multipoint plots is the partial flexibility of the native state of Beta3s (20). Its partially exposed nonpolar side chains that have been mutated in this work are involved in less-specific interactions with the rest of the peptide than buried side chains in the hydrophobic core of larger proteins. Because of the partial flexibility, the mutations do not affect only specific interactions but produce an effect that is spread over the large available set of contacts and thus averaged over them. This averaging of the effects of mutations in the native state may translate into a simple linear dependence of the effects in the TS. In this context, deviations from linearity may indicate TS shifts (see *Heterogeneity of the TS Ensemble*).

In multipoint plots, different local probes of the same residue are forced in a single fit that can yield wrong estimates (33). As an example, in the I → V → A → G mutation series. the I → V measures interactions originating from tertiary structure contacts, the V → A measures a mixture of tertiary and secondary structure interactions, whereas the A → G reports almost exclusively on secondary structure formation (33).

In a framework (34) or diffusion-collision (35) mechanism of folding, the tertiary Φ values will most probably be lower than secondary Φ values, even for the same residue. In the case of Beta3s, where the formation of β-sheet backbone HBs and long-range contacts between side chains are concomitant events (see figure 4 in ref. 22), different mutations probe the formation of the same level of structure (i.e., the β-sheet) with no distinction between secondary and tertiary components. This result supports the validity of the multipoint analysis for Beta3s that we do not want to generalize to proteins with more complex folds.

Given the peculiarities of Beta3s, i.e., concomitant formation of secondary and tertiary structure and partial flexibility of its folded

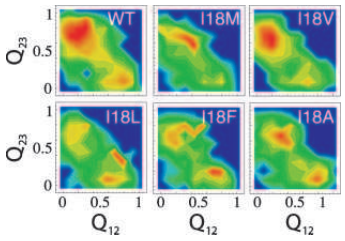

Fig. 4. Distribution of the fraction of native contacts in the N-terminal β-hairpin (Q_{12}) and C-terminal β-hairpin (Q_{23}) for the TS ensemble in the wild-type and I18 mutants. The color indicates the density of conformations and it changes from blue to red as the density increases. The two separated maxima correspond to the two possible folding pathways. (*Upper*) The more stable species. (*Lower*) The less stable species. A destabilization of >0.3 kcal/mol for I18 mutants results in a shift of the TS from β-hairpin 2–3 to β-hairpin 1–2.

state, multipoint Φ values may add information on the accuracy of the two-point Φ values. Indeed, reliable and tolerable Φ values fall mostly within an SD from the corresponding multipoint Φ value (Fig. 2), whereas unreliable Φ values show large deviations. Five of the six multipoint Φ values of Beta3s are >0.5. For diffuse TS ensembles of proteins of ≈100 residues, Φ values of ≈0.2–0.3 have been measured experimentally (36, 37). The high Φ values of Beta3s are probably because of the small size of the peptide. Because of its small size, a large part of the native interactions of the hydrophobic residues is already present in the rate-limiting step (see below).

Heterogeneity of the TS Ensemble. In the wild-type Beta3s, two parallel folding pathways were identified (22, 38). They correspond approximately to conformations having either of the two native β-hairpins formed and the remaining strand unstructured as revealed by the fraction of contacts formed in the two hairpins Q_{12} and Q_{23}. The TS conformations of the mutants have been analyzed and a similar scenario has been found in all of them. However, the relative abundance of the two pathways is different for different mutants. In most of the mutants, the most populated (thus, rate-limiting) pathway corresponds to the formation of the β-hairpin 2–3, followed by the formation of the β-hairpin 1–2, as in the wild type. In some of the mutants of I3, W10, and I18, the relative weight of the two pathways is inverted. In the multipoint plot of I18 (Fig. 3), the wild-type and the less destabilized mutants (i.e., I18V and I18M) lie on a much steeper line (slope = 1.8) than the more unstable I18F and I18A (slope = 0.2). I18L lies on the crossing of the two lines. The presence of a kink in the linear relationship in the multipoint plot indicates a shift in the folding pathway (24, 39), as confirmed by structural analysis of the TS ensemble of wild type and mutants of I18 (Fig. 4) Wild type, I18V, and I18M have a TS ensemble with β-hairpin 2–3 that is more structured than β-hairpin 1–2 (i.e., $Q_{23} > Q_{12}$). On the other hand, for the remaining mutants, the population of the pathways is either similar (I18A), or β-hairpin 1–2 is more structured than β-hairpin 2–3 (I18F and I18L), revealing a shift in the folding pathway determined by the destabilization of β-hairpin 2–3. This destabilization could be a consequence of different steric requirements of γ-branched side chains (Leu and Phe) with respect to β-branched (Val) or unbranched (Ala and Met). A similar shift is observed for the mutants of I3, where a destabilization >0.3 kcal/mol leads to a structural change of the TS (data not shown). Whereas for mutants of I3 and I18, the TS shift can be inferred from the multipoint plot, this is not the case for the W10L, W10Y, and W10V mutants, whose distribution of Q_{12} and Q_{23} at the TS (data not shown) indicates a more frequent folding pathway through early formation of β-hairpin 1–2.

The structural Φ values, i.e., the amount of contacts formed at the

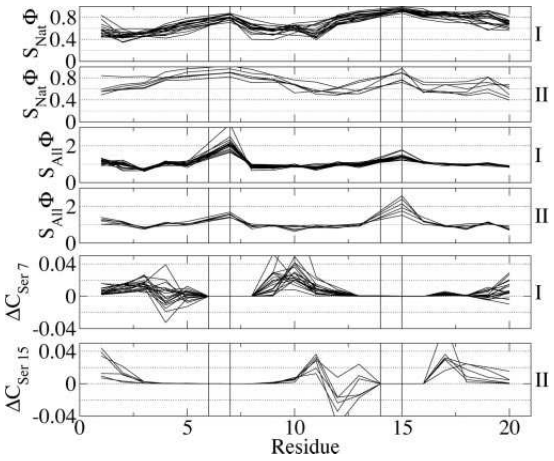

Fig. 5. Heterogeneity and nonnative structure of TS. The labels on the right indicate mutants with the C-terminal β-hairpin more structured at the TS (I) and mutants with the N-terminal β-hairpin more structured at the TS (II). Each solid curve corresponds to a single-point mutant and the lines are drawn to help the eye. Vertical lines indicate the position of the G6-S7 and G14-S15 turns. In the first four rows, the structural Φ values (S_{Nat} Φ and S_{All} Φ) are the ratio between the number of contacts formed in the TS and native state. S_{Nat} Φ takes into account only native contacts, whereas S_{All} Φ also includes nonnative contacts at the TS and can assume values >1. In the last two rows, $\Delta C_X(R) = \Sigma_{i \in C(X,R)} [p_{TS}(i) - p_N(i)]$ is the difference between the contacts formed in the TS and in the native state between residue X and R. Positive values indicate that in the TS, there are more contacts than in the native state. Both S7 and S15, if the corresponding hairpin is not native (i.e., in the class of mutants I and II, respectively), have a larger number of contacts in the TS than in the native state with K9 and the residue in position 10, and K17 and the residue in position 18, respectively. A smaller number of contacts in the TS than in the native state is observed with $Q4$ and $Q12$, respectively.

TS ensemble relative to the native state, provide a precise indication of the distribution of structure at TS with respect to the native state. The $S_x\Phi$ profiles can be divided in two major classes (Fig. 5). The first class (I) contains all of the mutants with a TS that is more structured around the C-terminal G14-S15 turn, according to the $S_{Nat}\Phi$ values, whereas the structure around the N-terminal turn features many nonnative interactions, according to the large $S_{All}\Phi$. This class contains the wild type, the mutants of W2, Y11 and Y19, and the mutants I3V, I3M, W10G, I18V, and I18M. The second class (II) contains the mutants that have a TS more structured around the N-terminal turn, as reported by the $S_{Nat}\Phi$ values, whereas the C-terminal turn is involved in many nonnative interactions, as shown by the $S_{All}\Phi$ profile. This class contains the mutants I3A, W10L, W10Y, W10V, I18F, and I18L. The remaining mutants show $S_x\Phi$ profiles that lie between the two major classes (data not shown).

Specific Nonnative Structure in the TS. The large number of nonnative interactions made by S7 and S15 in peptides of class I and II, respectively, at the TS (Fig. 5) is mainly constituted by contacts with the lysine residue in position $i + 2$ (K9 and K17) and with the residue in position $i + 3$. On the other hand, the contacts of S7 and S15 with $Q4$ and $Q12$, respectively, are significantly less in the TS than in the native state. The secondary structure analysis of the G6-S7/G14-S15 residues in the disordered hairpin at TS indicates them as forming a turn in most of the conformations. However, the HBs between residues N5 and T8 (N13 and T16), characterizing the native type II′ turn, are present only in 34% (40%) of the TS structures of the mutants of class I (II). Furthermore, no other

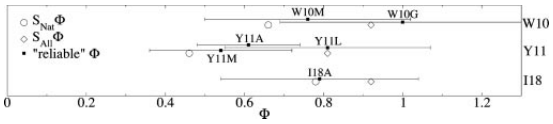

Fig. 6. Comparison between reliable two-point Φ values (filled squares) of mutants with a TS similar to wild type, and the structure of wild-type TS as measured by $S_x\Phi$ values (open symbols). The structural Φ values are the ratio between the number of contacts formed in the TS and native state. $S_{Nat}\Phi$ takes into account only native contacts, whereas $S_{All}\Phi$ includes native and nonnative contacts. The two-point Φ values tend to overestimate the degree of nativeness of the TS (measured by $S_{Nat}\Phi$) because of the presence of specific nonnative interactions.

specific backbone HBs are formed that define different types of turn. All these data indicate that the precursors of the type II′ turn, formed by the G-S pair of amino acids, are prevalently loose turns devoid of a specific backbone HB pattern that are shifted by one residue to the C terminus. Nonnative interactions, thus, are specifically involved in determining the commitment to fold of a conformation.

Structural Interpretation of Φ Values. Both $S_{Nat}\Phi$ and $S_{All}\Phi$ profiles of wild-type Beta3s provide a detailed picture of its TS. A comparison has been made with the reliable Φ values derived from mutations that do not change significantly the TS of the peptide (i.e., W10M, W10G, Y11A, Y11L, Y11M, and I18A), as indicated by the similarity of the $S_x\Phi$ profiles (namely, none of these mutants belong to class II, see above). This analysis allows for the assessment of the common interpretation of the Φ as a ratio between contacts formed at the TS and native states (Fig. 6). The comparison reveals that, within their error, the two-point Φ values are in agreement with both $S_x\Phi$s. However, the former tend to overestimate the degree of native structure present at the TS ensemble (i.e., reliable $\Phi > S_{Nat}\Phi$) because specific nonnative interactions are formed at the TS.

Conclusions

The near-equilibrium MD simulations of Beta3s and eight single-point mutants have provided an accurate estimate of Φ values for the mutations with stability changes of >0.6 kcal/mol. For such mutations, the SD on the value of Φ is relatively small, and the two-point Φ value is close to the corresponding multipoint Φ value, and to the structural Φ value that is a measure of the amount of contacts in the TS relative to the native state. In the other cases, the error is large and the estimate is less reliable. The value of the stability change threshold (0.6 kcal/mol) obtained from the simulation results of Beta3s and its mutants is smaller than the one proposed by Sanchez and Kiefhaber (1.7 kcal/mol) (23). Although it is not possible to extrapolate the simulation results to larger proteins with well defined hydrophobic cores, it is reassuring that the same validity threshold was suggested recently by Fersht and Sato (24) for Φ values of nondisruptive deletion mutations, and was used in a study of the CspB protein (40), whereas a very close threshold was used for the immunity proteins Im7 and Im9 (0.7 kcal/mol) (41).

The cluster P_{fold} progress variable has been used for the identification of TS structures. The TS ensemble of Beta3s and its single-point mutants is made up of two sets of conformations with either of the two β-hairpins folded. A TS shift from structured β-hairpin 2–3 to structured β-hairpin 1–2 has been observed for some of the mutants with different steric properties of the side chain, e.g., β-branched vs. γ-branched. Furthermore, the important role of specific nonnative interactions in the TS has been revealed. Indeed, when either of the two hairpins is formed in the TS, the residues corresponding to the native type II′ turn assume in the unstructured hairpin mainly the conformation of a loose turn shifted by one residue in the C-terminal direction. Specific nonnative contacts distinguish the TS conformations from other structures having the same native interactions but having different nonnative interactions. Hence, neglecting nonnative interactions may prevent a complete understanding of the factors that are responsible for protein folding.

We thank E. Guarnera and Dr. E. Paci for interesting discussions. The MD simulations were performed on the Matterhorn Beowulf cluster at the Informatikdienste of the University of Zurich. We also thank C. Bollinger, Dr. T. Steenbock, and Dr. A. Godknecht (University of Zürich, Zürich) for setting up and maintaining the cluster. This work was supported by the Swiss National Science Foundation.

1. Fersht, A. R., Matouschek, A. & Serrano, L. (1992) *J. Mol. Biol.* **224,** 771–782.
2. Fersht, A. R. (1999) *Structure and Mechanism in Protein Science: Guide to Enzyme Catalysis and Protein Folding* (Freeman, New York).
3. Daggett, V. & Fersht, A. (2003) *Nat. Rev. Mol. Cell Biol.* **4,** 497–502.
4. Li, A. J. & Daggett, V. (1996) *J. Mol. Biol.* **257,** 412–429.
5. Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. (2001) *Nature* **409,** 641–645.
6. Li, L. & Shakhnovich, E. I. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13014–13018.
7. Paci, E., Vendruscolo, M., Dobson, C. M. & Karplus, M. (2002) *J. Mol. Biol.* **324,** 151–163.
8. Gsponer, J. & Caflisch, A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6719–6724.
9. Lindorff-Larsen, K., Vendruscolo, M., Paci, E. & Dobson, C. M. (2004) *Nat. Struct. Mol. Biol.* **11,** 443–449.
10. Settanni, G., Gsponer, J. & Caflisch, A. (2004) *Biophys. J.* **86,** 1691–1701.
11. Cho, J. H., Sato, S. & Raleigh, D. P. (2004) *J. Mol. Biol.* **338,** 827–837.
12. Nauli, S., Kuhlman, B. & Baker, D. (2001) *Nat. Struct. Biol.* **8,** 602–605.
13. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003) *Nat. Struct. Biol.* **10,** 658–662.
14. Shimada, J. & Shakhnovich, E. I. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11175–11180.
15. Hubner, I. A., Shimada, J. & Shakhnovich, E. I. (2004) *J. Mol. Biol.* **336,** 745–761.
16. Du, R., Pande V. S., Grosberg, A. Y., Tanaka, T. & Shakhnovich, E. I. (1998) *J. Chem. Phys.* **108,** 334–350.
17. Chandler, D. (1978) *J. Chem. Phys.* **68,** 2959–2970.
18. Northrup, S. H., Pear, M. R., Lee, C. Y., McCammon, J. A. & Karplus, M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4035–4039.
19. Lenz, P., Zagrovic, B., Shapiro, J. & Pande, V. S. (2004) *J. Chem. Phys.* **120,** 6769–6778.
20. Rao, F. & Caflisch, A. (2004) *J. Mol. Biol.* **342,** 299–306.
21. De Alba, E., Santoro, J., Rico, M. & Jiménez, M. A. (1999) *Protein Sci.* **8,** 854–865.
22. Ferrara, P. & Caflisch, A. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10780–10785.
23. Sanchez, I. E. & Kiefhaber, T. (2003) *J. Mol. Biol.* **334,** 1077–1085.
24. Fersht, A. R. & Sato, S. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 7976–7981.
25. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
26. Ferrara, P., Apostolakis, J. & Caflisch, A. (2002) *Proteins* **46,** 24–33.
27. Cavalli, A., Ferrara, P. & Caflisch, A. (2002) *Proteins* **47,** 305–314.
28. Ferrara, P., Apostolakis, J. & Caflisch, A. (2000) *J. Phys. Chem. B* **104,** 5000–5010.
29. Eaton, W. A., Munoz, V., Hagen, S, J., Jas, G. S., Lapidus, L. J., Henry, E. R. & Hofrichter, J. (2000) *Annu. Rev. Biophys. Biomol. Struct.* **29,** 327–359.
30. Ferrara, P. & Caflisch, A. (2001) *J. Mol. Biol.* **306,** 837–850.
31. Hartigan, J. A. (1975) *Clustering Algorithms* (Wiley, New York).
32. Tou, J. T. & Gonzalez, R. C. (1974) *Pattern Recognition Principles* (Addison–Wesley, Reading, MA).
33. Fersht, A. R. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 14338–14342.
34. Baldwin, R. L. & Rose, G. D. (1999) *Trends Biochem. Sci.* **24,** 77–83.
35. Karplus, M. & Weaver, D. L. (1976) *Nature* **260,** 404–406.
36. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995) *J. Mol. Biol.* **254,** 260–288.
37. Daggett, V., Li, A. J., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1996) *J. Mol. Biol.* **257,** 430–440.
38. Davis, R., Dobson, C. M. & Vendruscolo, M. (2002) *J. Chem. Phys.* **117,** 9510–9517.
39. Warshel, A., Schweins, T. & Fothergill, M. (1994) *J. Am. Chem. Soc.* **116,** 8437–8442.
40. Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. (2004) *J. Mol. Biol.* **339,** 555–569.
41. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003) *J. Mol. Biol.* **326,** 293–305.

Chapter 6

Replica exchange molecular dynamics simulations of reversible folding (JCP (2003) 119, 4035)

Replica exchange molecular dynamics simulations of reversible folding

Francesco Rao and Amedeo Caflisch[a)]
Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

(Received 28 April 2003; accepted 21 May 2003)

The replica exchange molecular dynamics (REMD) approach is applied to a 20-residue three-stranded antiparallel β-sheet peptide. At physiologically relevant temperature REMD samples conformational space much more efficiently than constant temperature molecular dynamics (MD) and allows reversible folding (312 folding events during a total simulation time of 32 μs). The energetic and structural properties during the folding process are similar in REMD and conventional MD at the temperature values where there is enough statistics for the latter. The simulation results indicate that the unfolded state contains a significant amount of non-native interactions especially at low temperature. The folding events consist of a gradual replacement of non-native contacts with native ones which is coupled with an almost monotonic decrease of the REMD temperature. © *2003 American Institute of Physics.* [DOI: 10.1063/1.1591721]

I. INTRODUCTION

To accurately describe the thermodynamics and kinetics of complex systems, such as biological macromolecules, a thorough sampling of the relevant conformations is required. Since such systems have energetic and entropic barriers that are higher than the thermal energy at physiological temperature standard molecular dynamics (MD) techniques often fail to adequately sample the conformational space. A number of approaches to enhance sampling of phase space have been introduced.[1,2] They are based on multiple time steps,[3] modified Hamiltonians,[4–6] or generalized ensembles e.g., entropic sampling, multicanonical methods, replica exchange methods (REM).[7] REM is an efficient way to simulate complex systems at low temperature and is the simplest and most general form of simulated tempering.[8] Sugita and Okamoto have extended the original formulation into an MD based version (REMD) and tested it on the pentapeptide Met-enkephalin *in vacuo*.[9] Sanbonmatsu and Garcia have applied REMD to investigate the structure of Met-enkephalin in explicit water[10] and the α-helical stabilization by the arginine side chain which was found to originate from the shielding of main chain hydrogen bonds.[11] Furthermore, the energy landscape of the C-terminal β-hairpin of protein G in explicit water has been investigated by REMD.[12,13] Recently, a multiplexed approach with multiple replicas for each temperature level has been applied to large-scale distributed computing for the folding of a 23-residue miniprotein.[14] Starting from a completely extended chain, conformations close to the NMR structures were reached in about 100 trajectories (out of a total of 4000) but no evidence of reversible folding (i.e., several folding and unfolding events in the same trajectory) was presented.[14]

Even for a small protein it is currently not yet feasible to simulate reversible folding with a high-resolution approach, e.g., MD simulations with an all-atom model. The practical difficulties in performing such brute force simulations have led to several types of computational approaches and/or approximative models to study protein folding. An interesting approach is to unfold starting from the native structure[15,16] but a detailed comparison with experiments[17] is mandatory to make sure that the high temperature sampling does not introduce artefacts. Another possibility is offered by very small protein fragments for which the conformational space is sufficiently small so that full searches can be accomplished and/or transitions of interest occur on a manageable time scale.[18] The thermodynamic properties of two peptides (an α-helix and a β-hairpin of 13 and 12 residues, respectively) have been determined using an implicit solvation model and adaptive umbrella sampling.[19] Furthermore, the free energy surface of Betanova, an antiparallel three-stranded β-sheet peptide, has been constructed starting from conformations obtained during unfolding simulations in explicit water at elevated temperatures (between 350 and 400 K).[20]

In previous studies we have shown that it is possible to simulate the reversible folding of structured peptides at relatively high temperature values (330–360 K)[21–25] using an implicit model of the solvent based on the accessible surface area.[26] In this work we use REMD to explore, at temperature values of 275–465 K, the conformational space of Beta3s a designed peptide (Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17}-Ile_{18}-Tyr_{19}-Thr_{20}) whose solution conformation has been studied by NMR.[27] Nuclear Overhauser enhancement (NOE) and chemical shift data indicate that at 10 °C Beta3s populates a single structured form, the expected three-stranded antiparallel β-sheet conformation with turns at Gly6-Ser7 and Gly14-Ser15, in equilibrium with the random coil. Furthermore, Beta3s was shown to be monomeric in aqueous solution by equilibrium sedimentation and NMR dilution experiments.[27] The folding behavior and energy surface of Beta3s are investigated here using the same implicit model of the solvent[26] and a comparison is made with previous constant temperature MD simulations.[25] This approximation is justified by explicit water MD studies which have shown

[a)]Author to whom correspondence should be addressed. Phone: (41 1) 635 55 21; fax: (41 1) 635 68 62; electronic mail: caflisch@bioc.unizh.ch

that the solvent does not play a detailed role in the folding of Betanova.[20]

The present study was motivated by two main questions: Does REMD allow a thorough sampling of the relevant conformations of a three-stranded antiparallel β-sheet peptide at physiologically relevant temperatures? Do the energetic and structural properties during the folding events sampled by REMD correspond to those observed in constant temperature MD simulations? The simulation results indicate that both questions can be answered affirmatively.

II. METHODS

A. Model

The MD simulations and part of the analysis of the trajectories were performed with the CHARMM program.[28] The peptide was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 potential function[28,29]). The remaining hydrogen atoms are considered as part of the carbon atoms to which they are covalently bound (an extended atom approximation). The effective energy, whose negative gradient corresponds to the force used in the dynamics (see also below), is of the form

$$E(\mathbf{r}) = E_{vacuo}(\mathbf{r}) + G_{\mathrm{solv}}(\mathbf{r}); \tag{1}$$

for a molecule with N atoms at Cartesian coordinates $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$. The PARAM19 *vacuo* energy function is

$$E_{vacuo}(\mathbf{r}) = \frac{1}{2}\sum_{\text{bonds}} k_b(b-b_0)^2 + \frac{1}{2}\sum_{\substack{\text{bond}\\ \text{angles}}} k_\theta(\theta-\theta_0)^2 + \frac{1}{2}\sum_{\substack{\text{dihedral}\\ \text{angles}}} k_\phi[1+\cos(n\phi-\delta)]$$
$$+ \frac{1}{2}\sum_{\substack{\text{improper}\\ \text{dihedrals}}} k_\omega(\omega-\omega_0)^2 + \sum_{i>j} \varepsilon_{ij}^{\min}\left[\left(\frac{d_{ij}^{\min}}{r_{ij}}\right)^{12} - 2\left(\frac{d_{ij}^{\min}}{r_{ij}}\right)^{6}\right] + \sum_{i>j}\frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}},$$

where b is a bond length, θ a bond angle, ϕ a dihedral angle, ω an improper dihedral, r_{ij} is the distance between atoms i and j, q_i and q_j are partial charges, $d_{ij}^{\min}$ and $\varepsilon_{ij}^{\min}$ are the optimal van der Waals distance and energy, respectively, and $\epsilon(r_{ij})$ is a screening function. Parameters are given in Ref. 29.

An implicit model based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute.[26] In this approximation, the solvation free energy is given by

$$G_{\mathrm{solv}}(\mathbf{r}) = \sum_{i=1}^{M} \sigma_i A_i(\mathbf{r}), \tag{2}$$

for a molecule having M heavy atoms with Cartesian coordinates $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_M)$. $A_i(\mathbf{r})$ is the solvent-accessible surface computed by an approximate analytical expression[30] and using a 1.4 Å probe radius. The model contains only two surface-tension-like parameters: one for carbon and sulfur atoms ($\sigma_{\mathrm{C,S}} = 0.012$ kcal/mol Å^2), and one for nitrogen and oxygen atoms ($\sigma_{\mathrm{N,O}} = -0.060$ kcal/mol Å^2).[26] Furthermore, ionic side chains were neutralized[31] and a linear distance-dependent screening function [$\epsilon(r_{ij}) = 2r_{ij}$] was used for the electrostatic interactions. The CHARMM PARAM19 default cutoffs for long range interactions were used, i.e., a shift function[28] was employed with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parametrization of the force-field and implicit solvation model. The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent model have been used recently in MD simulations of folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues,[22–24] and small proteins of about 60 residues.[32,33] Despite the lack of friction due to the absence of explicit water molecules, the implicit solvent model yields a separation of time scales consistent with experimental data: helices fold in about 1 ns[21] (≈ 0.1 μs, experimentally[34]), β-hairpins in about 10 ns[21] (≈ 1 μs),[34] and triple-stranded β-sheets in about 100 ns[25] (≈ 10 μs).[27]

B. REMD simulations

The basic idea of REMD is to simulate different copies (*replicas*) of the system at the same time but at different temperature values. Each replica evolves independently by MD and every 1000 MD steps (2 ps), states i,j with neighbor temperatures are swapped (by velocity rescaling) with a probability $w_{ij} = \exp(-\Delta)$,[9] where $\Delta \equiv (\beta_i - \beta_j)(E_j - E_i)$, $\beta = 1/kT$ and E is the effective energy [Eq. (1)]. During the 1000 MD steps the Berendsen thermostat is used to keep the temperature close to a given value with a coupling of 0.1 ps. This rather tight coupling and the length of each MD segment (2 ps) allows the kinetic and potential energy of the system to relax. High temperature simulation segments facilitate the crossing of the energy barriers while the low temperature ones explore in detail the conformations present in the minimum energy basins. The result of this swapping between different temperatures is that high temperature replicas help the low temperature ones to jump across the energy barriers of the system.

In this study 8 replicas were used with temperatures (in K): 275, 296, 319, 344, 371, 400, 431, 465. The acceptance ratio of exchange between neighbor temperatures was around

TABLE I. Simulations performed.

#	Length (μs)	T (K)	Method	Starting conformation	Folding events	Folding time (μs)
2	8×1	275–465	REMD	native	152	0.064
2	8×1	275–465	REMD	unfolded	160	0.067
8	1	275	MD	unfolded	0	⋯
4	1	296	MD	unfolded	2	>0.44
4	2.7,2.7,2.8,4.4	330	MD	native	72	0.085
1	1	465	MD	unfolded	0	⋯

20% to 30%. Four REMD runs each with 8 replicas were performed: two started from the native structure and two from an unfolded structure obtained in a run at 330 K (see below). Each trajectory has a length of 1 μs for a total of 32 μs of simulation time (see Table I).

C. Constant temperature MD simulations

A series of control runs of 1 μs each were performed at the two lowest (275 K, 8 runs and 296 K, 4 runs) and the highest (465 K, 1 run) temperature value used in REMD. The Berendsen thermostat was used and the starting structure was the same unfolded conformation as the one employed in two of the four REMD runs. In addition, four simulations at the melting temperature of 330 K[25] started from the folded state (a total of 12.6 μs which were available from another study[35]) were used as a comparison for the folding process between conventional MD and REMD (see Table I).

For both REMD and constant temperature MD, the SHAKE algorithm[36] was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time step of 2 fs. Furthermore, the nonbonded interactions were updated every 10 dynamics steps and coordinate frames were saved every 20 ps for a total of 5×10^4 conformations/μs. A 1 μs run requires approximately 2 weeks on a 1.4 GHz Athlon processor and the REMD simulations were run in parallel on a Linux Beowulf cluster.

D. Native contacts and progress variables

The conformations sampled previously at 300 K were used to define a list of 26 native contacts, of which 11, 11, 2 and 2 involve residues in strands 1–2, 2–3, 2–2 and 3–3, respectively (see Table 2 of Ref. 23). These include 10 backbone hydrogen bonds (five on each β-hairpin) and 16 contacts between side chains. The folding progress variable Q_{nat} is defined as the fraction of contacts common to both the current conformation and the native structure. It can be plotted as a function of simulation time to monitor the amount of folded structure. For a given snapshot along a trajectory, a native hydrogen bond is considered formed if the O$\cdots$H distance is smaller than 2.6 Å. A native side chain contact is considered formed if the distance between geometrical centers is smaller than 6.7 Å. A conformation is considered folded when the fraction of native contacts Q_{nat} is larger than 0.85 ($Q_{\mathrm{nat}}>22/26$) and unfolded when it is smaller than 0.15 ($Q_{\mathrm{nat}}<4/26$).[23,25] The folding time is defined as the temporal interval between the first time point with $Q_{\mathrm{nat}}>22/26$ and the first time point with $Q_{\mathrm{nat}}<4/26$ just after the preceding folding event. The following subsets of native contacts were used for monitoring the folding pathways: Q_{1-2} is defined as the fraction of the 11 native contacts (5 hydrogen bonds and 6 side chain interactions) formed between strands 1 and 2, while Q_{2-3} as the fraction of the 11 native contacts between strands 2 and 3.[23]

The total number of contacts (N_{tot}) includes native and non-native ones, and is computed counting all hydrogen bonds and side-chain contacts between all pairs of residues at least three positions apart in the sequence. In addition, the contacts between side chains of the pairs of residues 8–10, 16–18 and 18–20 are included because they are considered native contacts. The fraction of total contacts Q_{tot} is $N_{\mathrm{tot}}/26$ where the denominator was chosen to facilitate the comparison with Q_{nat} (note that Q_{tot} can be larger than 1 but this happens sporadically).

E. Effective energy and free energy

The effective energy and free energy surfaces, determined by simulations and experiments, play an important role for the understanding of the protein folding reaction.[37] The effective energy is the sum of the intramolecular energy (CHARMM PARAM19 force field energy) and the solvation free energy [Eq. (1)]. The latter is approximated by the solvent accessible surface term[26] [Eq. (2)] and contains the free energy contribution of the solvent within the approximations of an implicit model of the water molecules. The effective energy does not include the configurational entropy of the peptide which consists of conformational and vibrational entropy contributions.[31] For a system in thermodynamic equilibrium, the difference in free energy in going from state A to state B is proportional to the natural logarithm of the quotient of the probability of finding the system in state A divided by the probability of state B.

Due to the complexity of the protein folding process, it is necessary to group states and project the energy onto one or two order parameters that characterize the system. The value of the effective energy is averaged within a bin defined by discretizing the reduced space. For the free energy profiles the probability of finding the system in a given bin is assumed to be proportional to the number of MD snapshots belonging to that bin. An arbitrarily chosen reference point (e.g., the fully unfolded state) is used as the denominator of the probability quotient.

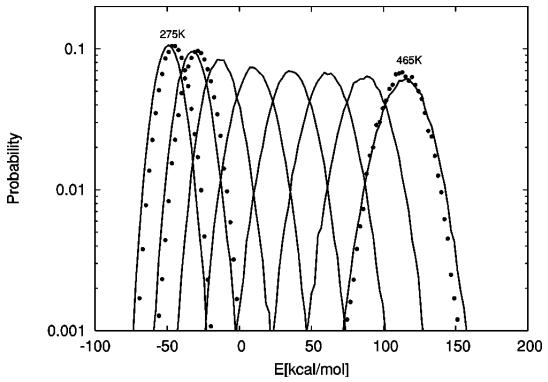

FIG. 1. Probability distribution of the effective energy in one of the two REMD runs starting from unfolded (solid lines) and constant temperature MD runs (filled circles). The REMD distributions correspond to the following temperatures (from left to right): 275, 296, 319, 344, 371, 400, 431, and 465 K. The constant temperature MD distributions (from left to right) correspond to 8 μs at 275 K, 4 μs at 296 K and 1 μs at 465 K.

III. RESULTS AND DISCUSSION

A. REMD diagnostics

The temperature values used in a REMD simulation have to be chosen carefully for an efficient sampling of the properties of interest. The highest value has to be high enough to jump over the energy barriers of the system, while the lowest value has to allow the exploration of the details of the energy minima. On the other hand, given a fixed number of replicas the range of temperature values cannot be too large because the acceptance probability for a temperature exchange has to allow a reasonable number of swaps during a simulation run. This implies that the temperature values need to be close enough to each other to guarantee a good overlap of the energy histograms (Fig. 1). An optimal distribution implies that the acceptance ratio for a swap between neighbor temperatures is nearly constant, resulting in a free random walk in temperature space. The filled circles in Fig. 1 show the results from the constant temperature MD simulations (at 275, 296, and 465 K) which were started from the same unfolded conformation used as a starting structure in two of the four REMD runs. The distributions agree at high temperature but tend to shift towards less favorable energies at low temperature values. This shows that conventional MD at low temperature can get trapped in local energy minima while REMD is superior in sampling conformational space. The time series of the exchanges of temperature values for one replica is shown in Fig. 2. It is clear that the trajectory visits all temperature levels several times within the 1 μs of simulation time. The other replicas show similar exchange patterns in all of the four REMD runs.

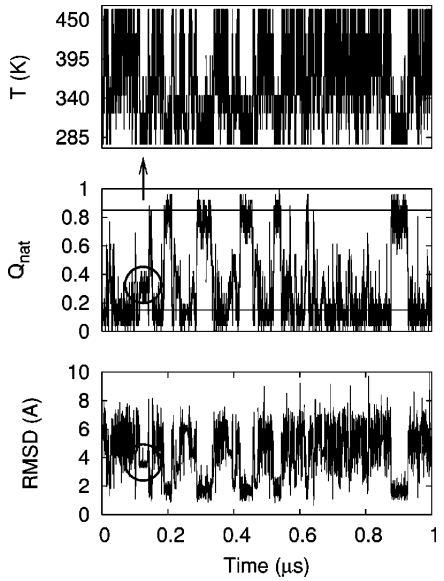

FIG. 2. Time series of (from top to bottom) the temperature T, fraction of native contacts Q_{nat}, and the C_α RMSD from the folded structure for a replica of a REMD run started from unfolded. The circles and arrow denote a simulation interval where the temperature is low but the peptide is not folded (see the text).

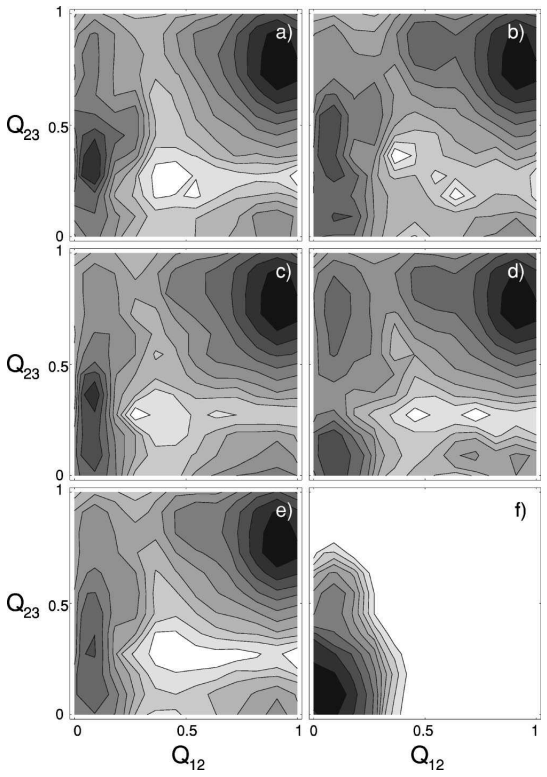

FIG. 3. Projection of the free energy surface onto the progress variables Q_{1-2} and Q_{2-3} at 275 K. (a)–(b) Two REMD runs started from unfolded. (c)–(d) Two REMD runs started from folded. In (a)–(d) each surface is plotted using 1 μs of data (5×10^4 conformations) sampled at 275 K. (e) Average over the four REMD runs. (f) Average over the eight 1-μs constant temperature MD runs at 275 K.

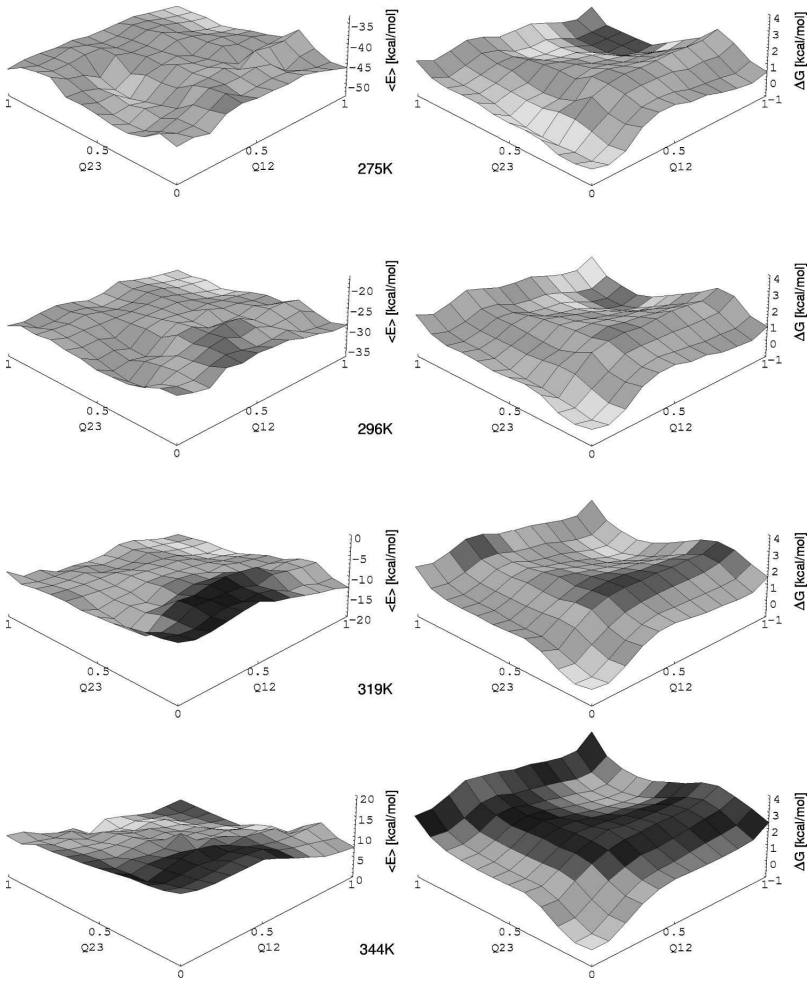

FIG. 4. Average effective energy (left) and free energy surface (right) as a function of the progress variables Q_{1-2} and Q_{2-3} in the REMD runs. The plots correspond to the following temperatures (from top to bottom) 275, 296, 319, and 344 K. A total of 2×10^5 conformations sampled during the 4 REMD runs were used for each value of the temperature.

B. Reversible folding

The time series of the fraction of native contacts (Q_{nat}) and C_α RMSD from the average NMR conformation indicate that several folding events are sampled along the REMD trajectories (Fig. 2). The C_α RMSD from native averaged over the time intervals where the structure is folded (e.g., 0.88–0.92 μs for the replica shown in Fig. 2) is 1.7 Å. A total of 312 folding events are sampled along the total simulation time of 32 μs. This corresponds to an average folding time of 0.065 ± 0.006 μs which is about 20% shorter than the value obtained by averaging over the 72 folding events sampled at 330 K constant temperature MD. It is important to note that there were only two folding events in the four 1 μs constant temperature runs at 296 K from unfolded (Table I). Moreover, in the eight 1 μs constant temperature runs at 275 K the value of Q_{nat} never exceeded 0.4 and the C_α RMSD from native was always above 3.5 Å.

Interestingly, in the REMD simulation intervals where the temperature is moderate (below 320 K) the peptide is folded most of the time but can also assume non-native conformations with values of $Q_{\mathrm{nat}}<0.4$ and RMSD>3.5 Å. This indicates that the unfolded state is explored at all temperature values (see the circles and arrow in Fig. 2).

C. Energy surfaces

Figure 3 shows a projection of the free energy surface for the REMD conformations sampled at 275 K and the constant temperature MD runs. It is clear that while the latter explores only a small portion of the unfolded state, the former samples both the folded and unfolded states. Moreover, if one neglects the noise due to frustration in the unfolded state the REMD surfaces at 275 K look similar among each other although two different initial conformations were used in the four REMD runs [Figs. 3(a)–3(d)]. This indicates that the choice of the initial conformation does not affect the REMD sampling.

Figures 4 and 5 show the average effective energy and free energy surfaces at the four low temperature values used in REMD. Raising the temperature results in a less rugged

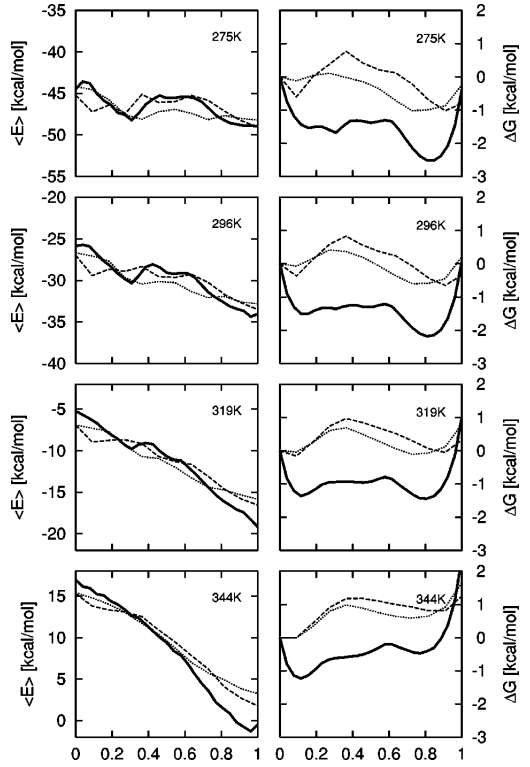

FIG. 5. Average effective energy (left) and free energy surface (right) as a function of the fraction of native contacts Q_{nat} (thick lines), Q_{1-2} (dashed lines), and Q_{2-3} (dotted lines). The same temperature values as in Fig. 4.

and more downhill profile of the effective energy. At 275 K and 296 K there is a pronounced minimum of the effective energy with values of Q_{2-3} in the range 0.3–0.7 and Q_{1-2} <0.3. This minimum was not observed in our previous constant temperature simulations and indicates that the β-hairpin 2–3 has an intrinsic enthalpic stabilization due to the interactions between the side chains of Trp10 and Tyr19 and Trp10 with the nonpolar part of the Lys17 side chain. This is consistent with an NMR analysis of two shorter peptides encompassing the single β-hairpins, i.e., peptides TWIQNGSTKWYQ and KWYQNGSTKIYT corresponding to Beta3s residues 1–12 and 9–20, respectively.[27] The NMR data at 10 °C indicate that the latter is as stable as Beta3s while the former is less stable.

The projection of the effective energy and free energy into Q_{1-2} and Q_{2-3} are consistent with our previous simulation results at constant temperature (330 K[25] and 360 K[23]). The free energy surfaces are more symmetric at higher temperature values (Fig. 4, right) because the entropic contribution starts to dominate and the conformational entropy penalty during folding is similar for both hairpins. The projections of the free energy along Q_{1-2} and Q_{2-3} (Fig. 5, right, dashed and dotted lines, respectively) indicate that the barrier is higher for the former especially at 275 and 296 K where the enthalpic contribution is more favorable for the formation of the second hairpin (Fig. 5, left). This effect is still present though less pronounced at higher temperatures and is consistent with previous constant temperature MD simulations at 330 K[25] and 360 K.[23]

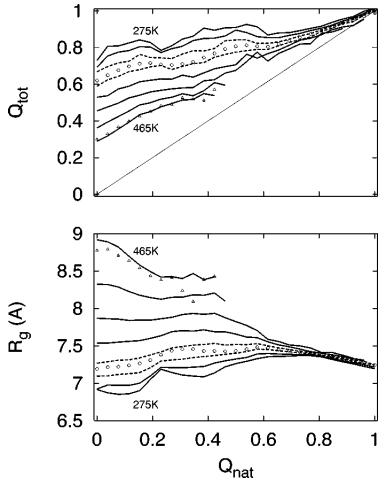

FIG. 6. Average fraction of total contacts (top) and average value of the radius of gyration (bottom) as a function of the fraction of native contacts Q_{nat}. REMD data (solid lines for all temperature values except for 319 K and 344 K in dashed lines) were obtained from the four runs, i.e., 4 μs for each temperature value. Constant temperature MD data are shown with symbols (12.6 μs at 330 K, circles; 1 μs at 465 K, triangles).

D. Non-native contacts and radius of gyration

At low temperature, non-native contacts in the unfolded state can act as traps. During folding, the total number of contacts grows moderately and non-native contacts are replaced by native ones [Fig. 6(a)]. At high temperature, the unfolded state is less compact [Fig. 6(b)] and has few non-native interactions so that the number of contacts during folding shows a more pronounced increase than at low temperature. From Fig. 6 it is evident that the behavior of total number of contacts and radius of gyration during folding is essentially the same in REMD and constant temperature MD at both medium (i.e., close to the melting temperature) and high temperature values. This provides additional evidence that the sampling of conformational space in REMD is correct.

E. Folding events

The previous analysis was based on the projections of energetic and structural properties along progress variables defined by the number of contacts present in the folded state or subsets thereof. Since folding is a complicated process with several degrees of freedom involved the choice of an adequate progress variable is not straightforward.[38] For this reason, it is useful to supplement the previous projections with the analysis of the folding mechanism during the time interval before reaching the folded structure. Figure 7 shows the behavior of temperature, number of native contacts and total number of contacts averaged over the 75 folding events that took longer than 5 ns along one of the two REMD runs

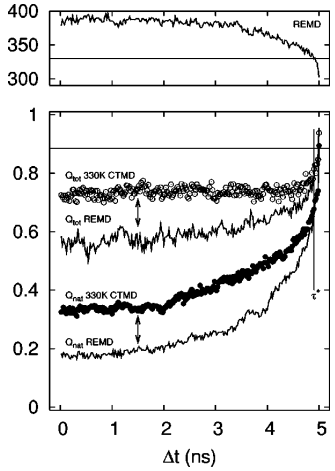

FIG. 7. The temperature (top) and average fraction of contacts (bottom) plotted during the 5 ns before reaching the folded structure for the REMD runs (solid lines) and the 330 K constant temperature MD simulations (circles). The REMD data are averages over 75 folding events along one of the two REMD runs started from unfolded. The constant line in the top marks the melting temperature of 330 K while the one in the bottom the definition of a folded structure, i.e., $Q_{\text{nat}} > 22/26$. The plot shows that the discrepancy in the total number of contacts between REMD and constant temperature MD simulations originates from the difference in native contacts (arrows) which implies that the number of non-native contacts is approximately the same. The time point τ^* at which the REMD temperature approaches 330 K is marked by a vertical line (bottom).

started from unfolded. In the last 5 ns before reaching the folded structure (defined by $Q_{\text{nat}} > 22/26$), the temperature on average decreases almost monotonically and the decrease is more pronounced the closer the system approaches the folded state. The conformations sampled during the folding process are characterized by minor variations in the number of contacts, i.e., an almost monotonic replacement of non-native interactions with native ones. At the beginning of the 5 ns time interval before reaching the folded structure the system is on average at elevated temperature in REMD with a value of $Q_{\text{nat}} \sim 0.18$ which is smaller than in the 330 K constant temperature MD ($Q_{\text{nat}} \sim 0.34$) because less contacts are formed at high temperature in REMD between strands 2 and 3 (not shown). Interestingly, the difference in the number of native contacts is the cause of the difference in the Q_{tot} curves (the arrows in Fig. 7) which shows that the two simulation types yield the same amount of non-native contacts during the folding process. This indicates that the REMD approach does not produce an artificial Go-like dynamics where non-native contacts are explicitly penalized. The REMD curves approach the same value of the constant temperature ones at the time point where the temperature is about 330 K (τ^* in Fig. 7).

IV. CONCLUSIONS

Four main results emerge from the REMD simulations of Beta3s with an implicit solvent. First, it is possible to sample the reversible folding of a structured peptide of 20 residues at physiologically relevant temperatures. This allows us to extract equilibrium properties even at low temperature and yields an atomic level description of the most populated conformations which is not (yet) feasible with conventional MD. The REMD approach is useful for an efficient sampling of the phase space and unlike other methods (like entropic sampling or the multicanonical method) REMD does not require the evaluation of the density of the states in a not obvious and tedious iterative procedure. Moreover, it can be implemented in a straightforward way on a parallel computer giving a scalability almost linear in the number of replicas used. Second, the important energetic and structural properties (e.g., average effective energy, number of non-native contacts, radius of gyration) monitored along the folding process are the same in REMD and constant temperature MD. The discrepancies at low temperature values are due to the limitations in sampling by constant temperature MD. Third, the effective energy surface at low temperature is more rugged and less symmetric with respect to the formation of the two hairpins than at high temperature. A similar but less pronounced temperature dependence is observed for the free energy surface. Fourth, the unfolded state can be investigated under folding conditions, namely physiologically temperatures, and contains a significant portion of non-native structures whose amount is inversely related to the temperature. The high amount of non-native interactions in the unfolded state at low temperature might be valid, in general, for structured peptides and will be analyzed in more detail by further simulation studies. In conclusion, REMD seems particularly useful to study the reversible folding of structured peptides (and probably small proteins in the near future) at the atomic level of detail.

ACKNOWLEDGMENTS

We are grateful to Dr. A. Cavalli, M. Cecchini, E. Paci, and G. Settanni for helpful discussions. We thank Dr. M. Seeber for developing software tools used to analyze the trajectories. We thank A. Widmer (Novartis Pharma, Basel) for providing the molecular modeling program Wit!P which was used for a visual analysis of the trajectories. The simulations were performed on a Beowulf cluster running Linux and we thank Urs Haberthür for his help in setting up the cluster. This work was supported by the Swiss National Competence Center in Structural Biology (NCCR) and the Swiss National Science Foundation (Grant No. 31-64968.01 to A. C.).

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic, San Diego, 2002).
[2] B. J. Berne and J. E. Straub, Curr. Opin. Struct. Biol. **7**, 181 (1997).
[3] T. Schlick, E. Barth, and M. Mandziuk, Annu. Rev. Biophys. Biomol. Struct. **26**, 181 (1997).
[4] X. Wu and S. Wang, J. Phys. Chem. B **102**, 7238 (1998).
[5] J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099 (1999).
[6] I. Andricioaei, A. R. Dinner, and M. Karplus, J. Chem. Phys. **118**, 1074 (2003).
[7] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[8] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[9] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[10] K. Sanbonmatsu and A. Garcia, Proteins: Struct., Funct., Genet. **46**, 225 (2002).
[11] A. E. Garcia and K. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).

[12] A. E. Garcia and K. Sanbonmatsu, Proteins: Struct., Funct., Genet. **42**, 345 (2001).
[13] R. Zhou, B. Berne, and R. Germain, Proc. Natl. Acad. Sci. U.S.A. **98**, 14931 (2001).
[14] Y. M. Rhee and V. S. Pande, Biophys. J. **84**, 775 (2003).
[15] A. Caflisch and M. Karplus, J. Mol. Biol. **252**, 672 (1995).
[16] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).
[17] U. Mayor, N. R. Guydosh, C. M. Johnson *et al.*, Nature (London) **421**, 863 (2003).
[18] D. Mohanty, R. Elber, D. Thirumalai, D. Beglov, and B. Roux, J. Mol. Biol. **272**, 423 (1997).
[19] M. Schaefer, C. Bartels, and M. Karplus, J. Mol. Biol. **284**, 835 (1998).
[20] B. D. Bursulaya and C. L. Brooks III, J. Am. Chem. Soc. **121**, 9947 (1999).
[21] P. Ferrara, J. Apostolakis, and A. Caflisch, J. Phys. Chem. B **104**, 5000 (2000).
[22] A. Hiltpold, P. Ferrara, J. Gsponer, and A. Caflisch, J. Phys. Chem. B **104**, 10080 (2000).
[23] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[24] P. Ferrara and A. Caflisch, J. Mol. Biol. **306**, 837 (2001).
[25] A. Cavalli, P. Ferrara, and A. Caflisch, Proteins: Struct., Funct., Genet. **47**, 305 (2002).
[26] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Genet. **46**, 24 (2002).
[27] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez, Protein Sci. **8**, 854 (1999).
[28] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[29] E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
[30] W. Hasel, T. F. Hendrickson, and W. C. Still, Tetrahedron Comput. Methodol. **1**, 103 (1988).
[31] T. Lazaridis and M. Karplus, Proteins: Struct., Funct., Genet. **35**, 133 (1999).
[32] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).
[33] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **99**, 6719 (2002).
[34] W. A. Eaton, V. Munoz, J. Hagen, S. G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, Am. Lab. (Boston) **29**, 327 (2000).
[35] A. Cavalli, U. Haberthür, E. Paci, and A. Caflisch, Protein Sci. (to be published).
[36] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[37] A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus, TIBS **25**, 331 (2000).
[38] A. Dinner and M. Karplus, J. Phys. Chem. B **103**, 7976 (1999).

Chapter 7

Replica exchange molecular dynamics simulations of amyloid peptide aggregation

(JCP (2004), 121, 10748)

JOURNAL OF CHEMICAL PHYSICS VOLUME 121, NUMBER 21 1 DECEMBER 2004

Replica exchange molecular dynamics simulations of amyloid peptide aggregation

M. Cecchini, F. Rao, M. Seeber, and A. Caflisch[a)]
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 6 May 2004; accepted 1 September 2004)

The replica exchange molecular dynamics (REMD) approach is applied to four oligomeric peptide systems. At physiologically relevant temperature values REMD samples conformation space and aggregation transitions more efficiently than constant temperature molecular dynamics (CTMD). During the aggregation process the energetic and structural properties are essentially the same in REMD and CTMD. A condensation stage toward disordered aggregates precedes the β-sheet formation. Two order parameters, borrowed from anisotropic fluid analysis, are used to monitor the aggregation process. The order parameters do not depend on the peptide sequence and length and therefore allow to compare the amyloidogenic propensity of different peptides. © *2004 American Institute of Physics*. [DOI: 10.1063/1.1809588]

I. INTRODUCTION

A thorough sampling of conformational space is required to describe the thermodynamics of complex systems such as multiple peptide chains at finite concentrations. Constant temperature molecular dynamics (CTMD) techniques often fail to adequately sample conformational space of frustrated and minimally frustrated systems which are characterized by a rugged free-energy landscape where energy barriers between minima are higher than the thermal energy at physiological temperature. For this reason, a number of approaches to enhance sampling of phase space have been introduced.[1–4] The parallel tempering technique (also known as replica exchange) was developed for dealing with the slow dynamics of disordered spin systems.[5] Sugita and Okamoto have extended the original formulation of replica exchange into an MD based version (REMD) and tested it on the pentapeptide Met-enkephalin *in vacuo*.[6] Although in the context of fragile liquids De Michele and Sciortino found that parallel tempering does not increase the speed of equilibration of the (slow) configurational degrees of freedom,[7] in the case of atomistic simulations of proteins many different applications have shown the efficiency of the method. Sanbonmatsu and Garcia have used REMD to investigate the structure of Met-enkephalin in explicit water,[8] and the α-helical stabilization by the arginine side-chain which was found to originate from the shielding of main chain hydrogen bonds.[9] REMD has also been applied to investigate the energy landscape of the C-terminal β-hairpin of protein G[10,11] and a three-helix bundle protein.[12] REMD in implicit solvent has been used to investigate the thermodynamics of designed 20-residue structured peptides,[13,14] and recently to study folding of a helical transmembrane protein.[15]

Highly ordered protein aggregates are associated with severe human disorders including Alzheimer's disease, type-II diabetes, systemic amyloidosis, and transmissible spongiform encephalopathies.[16,17] The soluble precursors of the ordered protein deposits do not share any sequence homology or common fold. However, x-ray diffraction data indicate a cross-β-structure for most fibrillar aggregates.[18,19] These findings suggest that key steps in the aggregation process may be common to all amyloidogenic proteins. Despite the medical relevance of amyloidoses, many important questions about the formation of ordered aggregates remain unanswered. There is experimental evidence that cytotoxicity is more pronounced for the early aggregates than for highly organized fibrillar structures.[20] Moreover, some peptide fragments of amyloidogenic proteins display the same properties as the full-length protein, including cooperative kinetics of aggregation, fibril formation, binding of the dye Congo red, and the cross-β x-ray diffraction pattern.[21] Both findings are particularly interesting because current simulation approaches allow significant sampling only for oligomeric peptide systems.

There have been several lattice studies on aggregation in proteins. These simplified models have allowed to investigate the foldability and aggregation propensity[22,23] and how interaction potentials affect the properties of aggregation-prone proteins.[24] Harrison *et al.* have shown that less stable proteins have a greater chance of assuming alternative native states as multimers.[25] MD simulations of aggregation have been performed by using a three-bead backbone and single-bead side-chain model.[26] While this simplified model has allowed the simulation of the competition between folding and aggregation for two four-helix bundles it is probably not possible to extract detailed information on energetics and sequence dependence. Recently, a minimalist Go model of four peptide strands[27] has been investigated by MD simulations in a confining sphere and the aggregation process was shown to depend on both sequence and environment.[28] Atomic models of amyloidogenic peptides have been simulated by MD with an implicit treatment of the solvent[29–31] and explicit water molecules.[32–36]

Recently, a replica exchange Monte Carlo technique has

[a)]Author to whom correspondence should be addressed. Fax: +41 1 635 68 62; Electronic mail: caflisch@bioc.unizh.ch

been applied to a lattice Go model of a minimalist multichain system to study the interplay between folding and disordered aggregation[23] but atomic model REMD applications to ordered aggregation have not been reported yet.

In the present paper, REMD with implicit solvent[37] is used to investigate the thermodynamics of the early steps of peptide aggregation and comparison is made with CTMD. The present work was motivated by three questions. Is it possible to sample the early events of ordered peptide aggregation at physiologically relevant temperatures? Do the aggregation energetics sampled by REMD correspond to those observed in CTMD simulations? Are the nematic and polar order parameters, borrowed from liquid crystal theory, useful to describe aggregation? The simulation results indicate that all questions can be answered affirmatively. Moreover, the "liquid crystal" order parameters allow to discriminate amyloidogenic peptide sequences from those that form only disordered aggregates.

II. METHODS

A. Model

The MD simulations and part of the analysis of the trajectories were performed with the CHARMM program.[38] The oligomeric peptide systems were modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 potential function[38,39]). The remaining hydrogen atoms are considered as part of the carbon atoms to which they are covalently bound (extended atom approximation). The effective energy, whose negative gradient corresponds to the force used in the dynamics, is

$$E(\mathbf{r}) = E_{vacuo}(\mathbf{r}) + G_{solv}(\mathbf{r}) \tag{1}$$

for a molecular system with atomic nuclei located at $\mathbf{r}=(\mathbf{r}_1,\ldots,\mathbf{r}_N)$. The PARAM19 *vacuo* energy function is

$$\begin{aligned} E_{vacuo}(\mathbf{r}) = {} & \frac{1}{2}\sum_{bonds} k_b(b-b_0)^2 + \frac{1}{2}\sum_{\substack{bond\\ angles}} k_\theta(\theta-\theta_0)^2 \\ & + \frac{1}{2}\sum_{\substack{dihedral\\ angles}} k_\phi[1+\cos(n\phi-\delta)] \\ & + \frac{1}{2}\sum_{\substack{improper\\ dihedrals}} k_\omega(\omega-\omega_0)^2 \\ & + \sum_{i>j} \varepsilon_{ij}^{\min}\left[\left(\frac{d_{ij}^{\min}}{r_{ij}}\right)^{12} - 2\left(\frac{d_{ij}^{\min}}{r_{ij}}\right)^{6}\right] \\ & + \sum_{i>j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}, \end{aligned}$$

where b is a bond length, θ a bond angle, ϕ a dihedral angle, ω an improper dihedral, r_{ij} is the distance between atoms i and j, q_i and q_j are partial charges, and $d_{ij}^{\min}$ and $\varepsilon_{ij}^{\min}$ are the optimal van der Waals distance and energy, respectively. Parameters are given in Ref. 39.

An implicit model based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute.[37] In this approximation, the solvation free energy is given by

$$G_{solv}(\mathbf{r}) = \sum_{i=1}^{N} \sigma_i A_i(\mathbf{r}) \tag{2}$$

for a molecular system having N heavy atoms with Cartesian coordinates $\mathbf{r}=(\mathbf{r}_1,\ldots,\mathbf{r}_N)$. $A_i(\mathbf{r})$ is the solvent-accessible surface computed by an approximate analytical expression[40] and using a 1.4 Å probe radius. The solvation model contains only two σ parameters: one for carbon and sulfur atoms ($\sigma_{C,S}=0.012$ kcal/mol Å^2), and one for nitrogen and oxygen atoms ($\sigma_{N,O}=-0.060$ kcal/mol Å^2).[37] Hence, according to Eq. (2) hydrophobic side chains tend to be buried within the solute whereas hydrophilic side chains and the polar groups of the backbone prefer to be solvent accessible. Furthermore, ionic side chains were neutralized[41] and a linear distance-dependent screening function [$\epsilon(r_{ij})=2r_{ij}$] was used for the electrostatic interactions. The CHARMM PARAM19 default cutoffs for long range interactions were used, i.e., a shift function[38] was employed with a cutoff at 7.5 Å for both the electrostatic and van der Waals terms. This cutoff length was chosen to be consistent with the parametrization of the force-field and implicit solvation model. The model is not biased toward any particular secondary structure type. In fact, exactly the same force field and implicit solvent model have been used recently in MD simulations of aggregation,[30,31] folding of structured peptides (α-helices and β-sheets) ranging in size from 15 to 31 residues,[42–44] and small proteins of about 60 residues.[45,46]

B. REMD simulations

The basic idea of REMD is to simulate different copies (*replicas*) of the system at the same time but at different temperatures values. Each replica evolves independently by MD and every t_{swap} states i,j with neighbor temperatures are swapped (by velocity rescaling) with a probability $w_{ij}=\exp(-\Delta)$,[6] where $\Delta\equiv(\beta_i-\beta_j)(E_j-E_i)$, $\beta=1/kT$, and E is the effective energy [potential and solvation energy, Eq. (1)]. A t_{swap} of 10 000 MD steps (20 ps) was chosen in order to allow the kinetic and potential energy of the system to relax. High temperature simulation segments facilitate the crossing of the energy barriers while the low temperature ones explore in detail energy minima. The result of this swapping between different temperatures is that high temperature replicas help the low temperature ones to jump across the energy barriers of the system.

In this study six replicas were used with temperatures (in kelvin) 275, 296, 319, 344, 371, and 400. This range corresponds to a subset of values used in a previous study of reversible peptide folding with the same force-field and solvation model.[14] The acceptance ratios of exchange between neighbor temperatures ranged between 15% and 24%. Each trajectory has a length of 2 μs for a total of 12 μs of simulation time (see Table I).

TABLE I. Simulations performed.

Peptide sequence	Length (μs)	T (K)	Method	IP aggregation events	IA aggregation events
GNNQQNY	10×0.5	275	CTMD	0	6 (19.2)[a]
GNNQQNY	5×1.0	296	CTMD	3 (14.4)	5 (1.6)
GNNQQNY	10×3.4	330	CTMD	54 (7.6)	43 (1.4)
GNNQQNY	2×1.0	371	CTMD	0	0
GNNQQNY	6×2.0	275–400	REMD	14 (60.3)	15 (3.9)
QQQQQQQ	6×2.0	275–400	REMD	27 (54.8)	2 (9.4)
AAAAAAA	6×1.0	275–400	REMD	4 (0.8)	12 (0.9)
SQNGNQQRG	6×2.0	275–400	REMD	1 (1.6)	6 (1.0)

[a]The average time (ns) the three peptides remained aggregated in IP and IA is given in parentheses.

C. Constant temperature MD simulations

A series of control runs were performed at constant temperature: (i) ten simulations at 330 K (total of 34 μs) used as a comparison for the aggregation process between CTMD and REMD (see Table I), (ii) ten 0.5 μs simulations at 275 K and (iii) five 1 μs simulations at 296 K to compare CTMD and REMD sampling at physiologically relevant conditions, and (iv) two 1 μs simulations at 371 K to study the system near the *condensation* temperature (see below).

For both REMD and CTMD, Langevin dynamics with a friction value of 0.15 ps^{-1} was used. This friction coefficient is much smaller than the one of water (43 ps^{-1} at 330 K computed as $3\pi\eta d/m$,[47] where η is the viscosity of water at 330 K, and d and m are the effective diameter, i.e., 2.8 Å, and mass of a water molecule, respectively) to allow for sufficient sampling within the μs time scale of the simulation. The small friction does not influence the thermodynamic properties of the system.

The SHAKE algorithm[48] was used to fix the length of the covalent bonds involving hydrogen atoms, which allows an integration time step of 2 fs. Furthermore, the nonbonded interactions were updated every ten dynamics steps and coordinate frames were saved every 20 ps for a total of 5×10^4 conformations/μs. A 1 μs run requires approximately two weeks on a 1.4 GHz Athlon processor and the REMD simulations were run in parallel on a Linux Beowulf cluster.

D. Progress variables

Aggregation contacts. In-register parallel and antiparallel aggregation contacts were defined following the prescription given in Ref. 30: a contact was considered to be present if the distance between two C_α atoms placed on different in-register strands was within 5.5 Å. The fraction of in-register parallel contacts Q_p and in-register antiparallel contacts Q_a were used to monitor the evolution of the aggregation process. In-register parallel and antiparallel aggregates, IP and IA, respectively, were considered formed when Q_p and Q_a were larger than 0.75 ($Q_p, Q_a > 11/14$) whereas at values smaller than 0.25 ($Q_p, Q_a < 4/14$), the system was considered disordered. The aggregation time is defined as the temporal interval between the first time point where $Q_p, Q_a < 0.25$ and the following time point where $Q_p, Q_a > 0.75$.

Radius of gyration. The radius of gyration of the oligomeric system R_g was considered to monitor the degree of *condensation* and calculated using the minimum image convention. Large values of R_g indicate conformations with isolated and non-interacting peptides (*uncondensed phase*). Small values of R_g indicate ordered as well as disordered aggregated conformations (*condensed phase*).

E. Orientational order parameters

The nematic and polar order parameters, $\overline{P_2}$ and $\overline{P_1}$, respectively, were considered in this study. These order parameters represent the first and second rank coefficients of the singlet orientational distribution expanded in a Wigner series,[49,50] i.e., a basis set of the Wigner rotation matrices. The nematic and polar order parameters are widely used for studying the properties of anisotropic fluids such as liquid crystals[51–54] and are defined as

$$\overline{P_2} = \frac{1}{N}\sum_{i=1}^{N} \frac{3}{2}(\hat{\mathbf{z}}_i \cdot \hat{\mathbf{d}})^2 - \frac{1}{2} \tag{3}$$

and

$$\overline{P_1} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{z}}_i \cdot \hat{\mathbf{d}}, \tag{4}$$

where $\hat{\mathbf{d}}$ (the director) is a unit vector defining the preferred direction of alignment, $\hat{\mathbf{z}}_i$ is a suitably defined molecular vector, and N is the number of molecules in the simulation box, i.e., three peptides in this study. The director is defined as the eigenvector of the ordering matrix,[55] that corresponds to the largest eigenvalue. Here, the molecular vectors $\hat{\mathbf{z}}_i$ were defined as unit vectors linking the peptide's termini (from the N to the C terminus, Fig. 1). To optimally select the $\hat{\mathbf{z}}_i$ vectors, other choices were investigated: vectors linking the carbonyl C to the amide N of each residue (''amide'' vectors) as well as vectors lying along the carbonyl bonds. Similar results were obtained with the three different choices of $\hat{\mathbf{z}}_i$. However, due to the atomic connectivity along the backbone the amide vectors are not fully independent; their orientations are strongly correlated and the description of the ordered macrostates results less precise. The same is true for the ''carbonyl'' vectors. Hence, vectors linking peptide's termini were preferred.

The order parameters [Eqs. (3) and (4)] change value on going from one order macrostate to the other and should vanish when the transition to a fully isotropic state takes place. They describe different orientational properties of the system and yield useful and complementary information. The

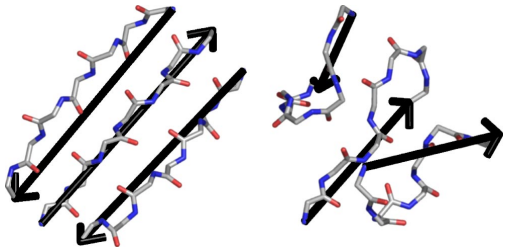

FIG. 1. Pictorial representation of the molecular vectors $\hat{\mathbf{z}}_i$ (black arrows) used to compute the order parameters $\overline{P_1}$ and $\overline{P_2}$. $\hat{\mathbf{z}}_i$ vectors are defined as full length peptide vectors (linking the peptide's termini) and allow to clearly discriminate between ordered (left, $\overline{P_2}=0.87$) and disordered (right, $\overline{P_2}=0.46$) conformations of the system. [The pictures were drawn using the program PYMOL (Ref. 66)].

nematic $\overline{P_2}$ describes the orientational order of the system and discriminates between ordered and disordered conformations. The polar $\overline{P_1}$ describes the polarity of the system, i.e., how much the molecular vectors $\hat{\mathbf{z}}_i$ point in the same direction, and discriminates between parallel and antiparallel/mixed ordered aggregates.

F. Peptides

To evaluate the reliability of amyloidogenic propensity estimations, four oligomeric peptide systems were considered in this study: the amyloid-forming heptapeptide GNNQQNY and the soluble nonapeptide SQNGNQQRG both from the yeast prion Sup35 (residues 7–13 and 17–25 with the Gln/Arg mutation at position 24, respectively),[21] the amyloidogenic poly(L-glutamine) QQQQQQQ (Ref. 56) and the nonamyloidogenic poly(L-alanine) AAAAAAA.[57] To reproduce the experimental conditions,[21,56,57] the peptide systems derived from the yeast prion Sup35 were modeled without blocking groups, while the Ala and Gln repeats were both *N*-acetylated and C-amidated.

All simulations were performed with three peptide replicas starting from random conformations, positions, and orientations. In the initial random positions there was no intermolecular contact, i.e., the peptides were separated in space. Each system was simulated in a cubic box of 75 Å per side yielding a sample concentration of 0.012 M. Since the oligomeric systems present different molecular weights, the above reported concentration corresponds to 3.4, 3.9, 5.4, and 3.4 mg/ml for GNNQQNY, SQNGNQQRG, QQQQQQQ, and AAAAAAA, respectively.

G. Analysis tools

The aggregation contacts, radius of gyration, and order parameters analysis was carried out with a GPL licensed program[58] developed in house to manipulate and analyze molecular dynamics (MD) trajectories. The program is optimized for speed and ease of usage so that it allows extensive processing of large amounts of data and straightforward addition of new analysis tools. Compared to other available programs,[38,59] the analysis of MD trajectories is much faster.

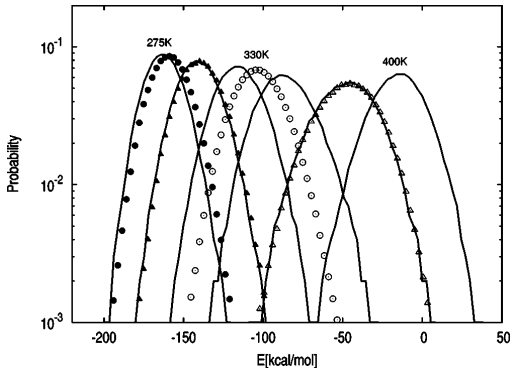

FIG. 2. Probability distribution of the effective energy for the REMD (solid lines) and the CTMD control simulations (filled circles, filled triangles, empty circles, and triangles for 275, 296, 330, and 371 K, respectively). The REMD distributions correspond to the following temperatures (from left to right): 275, 296, 319, 344, 371, and 400 K. The asymmetry of the curves and the temperature dependence of the distributions indicate the presence of a phase transition around 371 K (see text).

III. RESULTS AND DISCUSSION

A. REMD diagnostics

The set of temperatures used in a REMD simulation is crucial for a correct and efficient sampling.[8] Since a simple *a priori* protocol for selecting the optimal temperature distribution has not been identified (yet), the choice often follows empirical considerations:[8,14,23] the highest temperature of the set has to be high enough to overcome energy barriers, while the lowest temperature has to allow the exploration of minima. However, given a fixed number of replicas the temperature range cannot be too wide. Temperature values need to be close enough to make the energy histograms overlap (see Fig. 2) in order to guarantee a high number of temperature swaps during a simulation run. In this study, a set of six temperature values ranging from 275 to 400 K has been selected (see Methods). The time series of temperature exchanges for one of the six replicas is shown in Fig. 3. During the simulation, each replica visits all the temperatures of the set several times realizing the desired free random walk in temperature space.[6]

Symbols in Fig. 2 show the results from CTMD simulations carried out at 275 (filled circles), 296 (filled triangles), 330 (empty circles), and 371 K (empty triangles). At 330 K, the CTMD effective energy distribution is located between the REMD distributions extracted at 319 and 344 K and shows a consistent functional profile. At 371 K, CTMD and REMD effective energy distributions overlap. Therefore, the energetic properties of an aggregating system sampled by a REMD simulation at medium and high temperatures correspond to those observed in CTMD simulations. However, approaching the physiologically relevant conditions the CTMD distributions tend to shift toward less favorable energies (Fig. 2, filled symbols). CTMD at low temperature can get trapped in local energy minima and REMD is superior in sampling conformational space.[6,14]

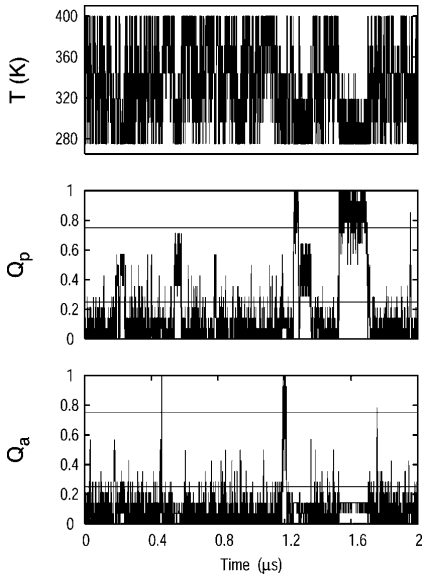

FIG. 3. Time series of (from top to bottom) the temperature T, the fraction of in-register parallel contacts Q_p, and the fraction of in-register antiparallel contacts Q_a for a REMD replica. Along the trajectory, replicas realize the desired free random walk in temperature space (top) so that an efficient sampling of the ordered aggregates is allowed (peaks in Q_p and Q_a plots). Horizontal lines in the time series of the fraction of aggregation contacts indicate the upper/lower thresholds used to define the ordered aggregation/disaggregation events.

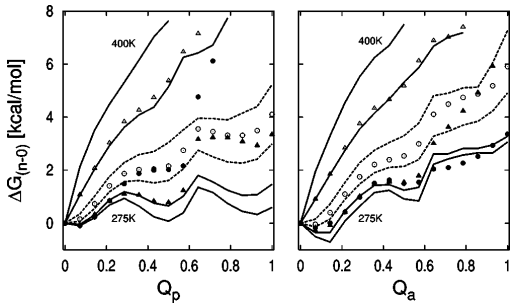

FIG. 4. Free-energy projections along the fraction of in-register parallel contacts Q_p (left) and in-register antiparallel contacts Q_a (right). Conformations with zero in-register contacts were chosen as reference states. $\Delta G_{(n-0)}$ was computed as $-k_B T \ln(N_n/N_0)$, where N_n indicates the number of conformations with n contacts and k_B is the Boltzmann constant. REMD data are shown in solid lines for all the temperature values except for 319 and 344 K which are in dashed lines. CTMD data are shown with symbols (filled circles, filled triangles, empty circles, and triangles for 275, 296, 330, and 371 K, respectively).

The time series of the fraction of in-register parallel contacts (Q_p) and in-register antiparallel contacts (Q_a) have been monitored along the REMD trajectories (Fig. 3). A total of 14 IP and 15 IA aggregation events have been observed along the total simulation time of 12 μs (see Table I). The average aggregation time (see Methods) was 0.74 μs for IP and 0.75 μs for IA arrangements. The average aggregation time determined from the REMD simulation is similar to the values obtained from 34 μs CTMD simulations at 330 K. It is worth noting that in a preliminary REMD run with higher temperatures values (6×1 μs, 319–465 K; data not shown) only 3 IP and 4 IA aggregation events were sampled. The temperature range is crucial in REMD and it has to be carefully chosen in order to speed up the conformational search of relevant states,[60] i.e., the *ordered states* when studying aggregation. To bias the search toward conditions where ordered states are more probable, the temperature was set to lower values (275–400 K, as mentioned above) and the sampling of aggregation events turned out substantially improved.

Figure 4 shows the projections of the free energy surface along Q_p and Q_a for both REMD and CTMD trajectories. The profiles indicate that the structural properties of the aggregating system sampled by a REMD simulation correspond to those observed in CTMD simulations only at high and medium temperatures. At 371 K, CTMD and REMD free energy projections overlap. At 330 K, the CTMD free-energy profiles (empty circles) are correctly placed between REMD projections at 319 and 344 K (dashed lines) and show patterns characterized by a well-defined local minimum at $Q_p > 0.7$ and a monotonic uphill trend along Q_a, fully consistent with the profiles extracted from the REMD simulation. However, at low temperature (275 and 296 K) the free energy profiles extracted from CTMD and REMD trajectories are not consistent any more and the most "relevant" conformations, which correspond to in-register parallel and antiparallel arrangements ($Q_p, Q_a > 0.7$), are not correctly sampled by CTMD (Fig. 4, filled symbols).

B. Temperature dependence of ordered amyloid peptide aggregation

Since the energetic and structural properties of the system are not artificially altered (see preceding section), the REMD approach allows to evaluate thermodynamic quantities as a function of temperature in the chosen range.[6] From the REMD simulation performed for this study, the properties of interest have been extracted at any temperature of the set and the aggregation of the amyloid-forming peptide GNNQQNY has been monitored in temperature space (275–400 K). This analysis gives interesting insights into the amyloid aggregation process.

The effective energy histograms shown in Fig. 2 are not symmetrically distributed around their mean value and their shape varies with temperature. The distributions, in fact, broaden toward higher energy values at low temperature (275–344 K) and toward lower energy values at high temperature (371–400 K). Moreover, by increasing temperature they progressively become lower and broader till the value of 371 K is reached. Mitsutake *et al.* have interpreted such a behavior as the evidence of a phase transition.[61] To characterize the transition, the radius of gyration R_g of the oligomeric system was considered and free-energy projections along R_g were plotted (see Fig. 5). Conformations of the system producing non-interacting peptides, namely, confor-

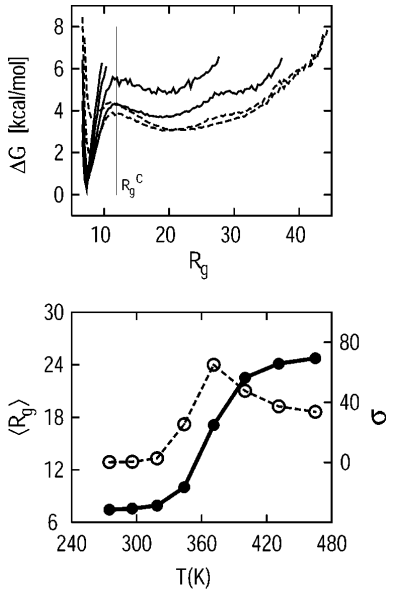

FIG. 5. (Top) Free-energy projections along the radius of gyration of the oligomeric system R_g computed from REMD trajectories. Solid lines correspond to temperature values below the condensation temperature (275–344 K); dashed lines correspond to temperature values above the transition temperature (371 and 400 K). The lowest radius of gyration for the uncondensed state is shown as a vertical line ($R_g^C = 11.9$ Å). (Bottom) Temperature dependence of the average radius of gyration $\langle R_g \rangle$ (filled circles) and its fluctuations σ (empty circles). The behavior of $\langle R_g \rangle$ and σ indicates the presence of a phase transition around 371 K between a condensed (low T) and an uncondensed phase (high T). Fluctuations of the radius of gyration σ are computed as $\langle R_g^2 \rangle - \langle R_g \rangle^2$. Data at 431 and 465 K were obtained from a preliminary REMD run carried out in a higher temperature range (6×1 μs, 319, 344, 371, 400, 431, and 465 K).

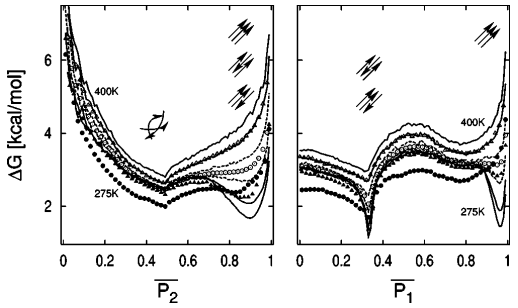

FIG. 6. Free-energy projections along the nematic ($\overline{P_2}$, left) and the polar ($\overline{P_1}$, right) order parameters. REMD data are shown in solid lines for all the temperature values except for 319 and 344 K which are in dashed lines. CTMD data are shown with symbols (filled circles, filled triangles, empty circles, and triangles for 275, 296, 330, and 371 K, respectively). Schematic representations of the aggregates (black arrows) are depicted to show that order parameters yield complementary information: $\overline{P_2}$ discriminates between ordered and disordered conformations while $\overline{P_1}$ discriminates between parallel and antiparallel/mixed ordered aggregates.

mations where all interpeptide atomic distances are larger than the long-range interactions cutoffs (7.5 Å in this case), were used to determine R_g^C, i.e., the lowest detected radius of gyration for isolated peptides (see Fig. 5). The existence of two macrostates in equilibrium has been revealed: the first, named *uncondensed state*, includes high energy conformations with one or more isolated peptides ($R_g > R_g^C$); the second, named *condensed state*, consists of low energy conformations with aggregated peptides ($R_g < R_g^C$). For entropic reasons, the *uncondensed state* is preferred at high temperature. By cooling down, the *condensed state* is increasingly stabilized, and around 371 K the fluctuations of R_g show a well-defined peak highlighting the presence of the *condensation* transition (see Fig. 5). The equilibrium between the *condensed* and the *uncondensed* macrostates is clearly concentration dependent. If the concentration of amyloid-forming units increases, the equilibrium is moved toward the condensed state and the aggregation process is favored.

The free-energy profiles along Q_p and Q_a at various temperatures help in understanding how the nucleation process evolves upon peptides condensation. At values of 400, 371, and 344 K both projections show steep uphill patterns with a single free-energy minimum at $Q_p \approx Q_a \approx 0$ (see Fig. 4). This means that upon condensation the peptides are still more likely to form disordered aggregates characterized by nonspecific interactions than amyloid-forming nuclei. In this range of temperatures, the enthalpic contribution due to in-register backbone or side-chain interactions does not dominate the entropic one and the growth of ordered nuclei is forbidden. However, when the temperature decreases the entropic contribution becomes less important and ordered in-register aggregates start forming. As shown in Fig. 4 in fact, below 330 K two and one additional free-energy minima appear in the projection along Q_p and Q_a, respectively. The observed minima correspond to in-register parallel ($Q_p > 0.7$) and in-register mixed or out-of-register ($0.3 \leqslant Q_p \leqslant 0.7$ and $0.4 \leqslant Q_a \leqslant 0.7$) arrangements and strongly suggest that the three-peptide system moves toward a higher degree of order when approaching the physiologically relevant conditions.

The simulation results indicate that in the early steps of amyloid aggregation a condensation stage toward disordered aggregates precedes the nucleation process and the disorder-order transition, in agreement with experimental evidence.[62]

C. Disorder-order transition

In the early steps of aggregation, amyloidogenic peptides assemble into highly ordered β-sheet structures.[21,30] During the assembly, the peptides tend to align adopting an extended β-strand conformation and a remarkable change in the local orientational order occurs. The aggregation of amyloid-forming peptides may then be interpreted as an order transition and orientational order parameters are suitable to monitor the time evolution of the process. Two orientational order parameters were employed and free-energy projections are shown in Fig. 6. Along $\overline{P_2}$, the free-energy profiles show a first broad minimum at $\overline{P_2} \approx 0.5$ for any temperature of the set and a second narrower one at $\overline{P_2} \approx 0.9$ for T values below

330 K. The first corresponds to a large free-energy basin where orientational order is absent, while the second corresponds to a smaller and well-defined basin with a high orientational degree of order. Although the order parameters should vanish when order is absent, Fig. 6 shows that this is not the case when the number of vectors is small. Since only three peptides were simulated, a "background" order was always detected and the free energy minimum describing the *disordered state* is placed at $\overline{P_2} \approx 0.5$, which is consistent with the value of $\sqrt{81/40\pi N}$ expected for a completely randomly oriented array of N molecules.[63] The order parameter $\overline{P_2}$ shows the existence of two macrostates in equilibrium: the disordered state with a high entropy content, which corresponds to the global minimum of the free energy surface at high temperature, and the *ordered state* which becomes the global free energy minimum at low temperature. Interestingly, the free-energy profiles along Q_p and Q_a do not lead to the same conclusion and the observed in-register arrangements correspond to local minima of the free-energy surface (see Fig. 4).

Along $\overline{P_1}$, two narrow and well-distinct minima corresponding to ordered macrostates at different polarity appear on the free-energy projections (Fig. 6). The first, displayed at $\overline{P_1} \approx 0.35$, describes a free-energy basin with a high-order and low-polarity content. Conversely, the second, displayed at $\overline{P_1} \approx 0.95$, corresponds to a basin with a high-order and high-polarity content. The order parameter $\overline{P_1}$ discriminates between parallel and antiparallel/mixed ordered conformations and provides complementary information since it allows to further characterize the ordered state.

Symbols in Fig. 6 show the free-energy projections along the order parameters from CTMD simulations. Once again, the comparison with REMD profiles indicates that isothermal MD (filled symbols) does not sample the ordered aggregates with their correct statistical weight close to the physiological temperature range.

The REMD free energy profiles along $\overline{P_1}$ show that at low temperature (275 and 296 K) both polar macrostates are highly populated. In the investigated temperature range, the system does not show an overall polar degree and frequent jumps between ordered states characterized by different polarity are observed. This suggests that below the order transition the equilibrium between polar macrostates might help amyloidogenic systems overcoming the entropy loss occurring during nucleation. In other words the growth of amyloid-forming nuclei might have an entropically favorable component due to the multiple ordered macrostates.

D. Sequence dependence of amyloidogenic propensity

Free-energy projections along the nematic order parameter $\overline{P_2}$ show how the equilibrium between the ordered and disordered state changes in temperature space (Fig. 6). Upon cooling, the statistical weight of the ordered state increases and the mean of the $\overline{P_2}$ distribution moves toward higher values. The value of $\langle \overline{P_2} \rangle$, where $\langle \cdots \rangle$ indicates the average over the canonical ensemble, is then related to the thermodynamic stability of the ordered state and could be used to measure the amyloidogenic propensity of the system. $\langle \overline{P_2} \rangle$ values computed at different temperatures from REMD trajectories of the amyloid-forming peptide GNNQQNY are shown in Fig. 7 with filled circles. At high temperature, the $\langle \overline{P_2} \rangle$ values are close to 0.5 because no orientational order is present, and the system does not show amyloidogenicity. By decreasing temperature, the amyloidogenic propensity grows and becomes increasingly larger until the order transition is completed. At physiologically relevant conditions, $\langle \overline{P_2} \rangle \approx 0.65$ and the system is highly amyloidogenic in agreement with experimental data.[21]

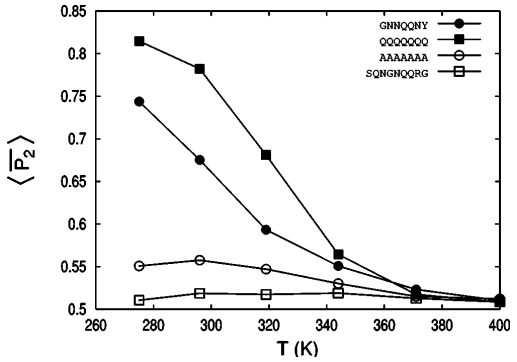

FIG. 7. Temperature dependence of the nematic order parameter $\langle \overline{P_2} \rangle$ averaged over the canonical ensembles sampled by REMD for four oligomeric peptide systems. $\langle \overline{P_2} \rangle$ estimates the amyloidogenic propensity of peptide systems and discriminates between amyloidogenic (GNNQQNY and QQQQQQQ) and nonamyloidogenic (SQNGNQQRG and AAAAAAA) sequences in agreement with experimental data (Refs. 21, 56, and 57).

Since the orientational order parameters do not depend on the peptide sequence and length, the reliability of the predictions could be further tested in sequence space. The REMD protocol was then applied to three additional oligomeric peptide systems (see Methods) and $\langle \overline{P_2} \rangle$ values were evaluated to measure and compare amyloidogenic propensities. The testing set comprises a nonapeptide from the yeast prion Sup35 (SQNGNQQRG) experimentally studied by Balbirnie *et al.*[21] and two heptapeptides (QQQQQQQ and AAAAAAA). Glutamine and alanine homopolymers flanked by basic residues to improve solubility have been investigated by Perutz *et al.*[56,57]

Experimentally, the nonapeptide SQNGNQQRG shows solubility *in vivo* and *in vitro* and no formation of amyloid fibrils.[21] In agreement with these findings, $\langle \overline{P_2} \rangle$ is smaller than 0.55 in the whole temperature range (Fig. 7, empty squares) and the system is considered as nonamyloidogenic. The number of aggregation events and the average lifetime of aggregation extracted from REMD trajectories are reported in Table I. Remarkably, these quantities show that nonamyloidogenic sequences, i.e, SQNGNQQRG and AAAAAAA, do transiently assemble in a β-sheet conformation but still remain soluble because their ordered aggregates do not correspond to well-defined free-energy minima.

Circular dichroism (CD) spectra, electron micrographs, and x-ray diffraction photographs showed that poly(L-

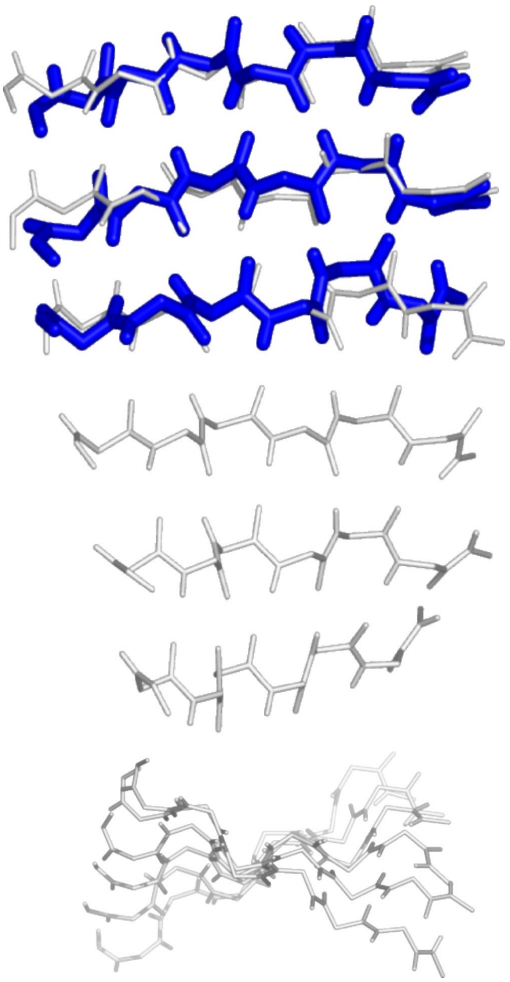

FIG. 8. (Top) Snapshots of ordered aggregates of three (thick sticks) and six (thin sticks) amyloidogenic SYVIIE peptides (Ref. 65) extracted from CTMD simulations at 330 K. The simulations were performed at a sample concentration of 5 mg/ml. The overall conformation and twist of the three-stranded and six-stranded parallel β-sheets are indistinguishable. (Bottom) The six-stranded β-sheet upon 90° rotation to better visualize the twist. [The pictures were drawn using the program PYMOL (Ref. 66)].

glutamine) peptides aggregate in solution at both pH 7.0 and 3.0 forming tightly linked β-sheet structures.[56] In particular, the x-ray diffraction picture exhibits a fiber diagram of the cross-β type distinctive of amyloid fibrils. On the other hand, poly(L-alanine) does not display amyloidogenicity and CD spectra showed α-helical structures at all pHs.[57] Again, the $\langle \overline{P_2} \rangle$ patterns shown in Fig. 7 (filled squares and empty circles) are consistent with experimental findings and correctly indicate amyloidogenicity only for QQQQQQQ.

Interestingly, Fig. 7 allows also to compare between amyloidogenic sequences. In fact, according to the $\langle \overline{P_2} \rangle$ patterns the glutamine repeat is more amyloidogenic than GNNQQNY at physiologically relevant conditions. To our knowledge, no experimental data are available to verify this finding. Testing of this prediction is a challenge for experimentalists.

IV. CONCLUSIONS

The present study shows that atomistic REMD simulations with implicit solvent allow to sample the early steps of ordered aggregation of amyloidogenic peptides at physiologically relevant temperatures. The free-energy profiles projected along structural and orientational progress variables are essentially the same in REMD and CTMD. The discrepancies at temperature values below 330 K are due to the limitations in sampling in CTMD simulations which indicates that REMD is a more efficient approach in the physiological range.

The early steps of amyloidosis can be interpreted as a condensation followed by an order transition. Therefore, the REMD simulation results were analyzed with two order parameters originally introduced to study liquid crystals. Interestingly, the nematic order parameter averaged over a canonical ensemble is able to discriminate amyloidogenic from soluble peptides in agreement with experimental data.

Although the present study was performed with three peptides for reasons of computational efficiency, the description of the ordered aggregates is likely to be independent of the size of system, i.e., the number of simulated peptide replicas. Very recent MD simulations of the amyloidogenic SYVIIE peptide,[64] which has been experimentally investigated by de la Paz and Serrano,[65] have shown ordered aggregates of six peptides. Interestingly, the parallel β-sheet consisting of six peptides has the same overall conformation and twist as the three-peptide aggregate (Fig. 8).

ACKNOWLEDGMENTS

We thank Dr. U. Haberthür for running most of the CTMD simulations and R. Pellarin for introducing periodic boundary conditions in the SASA module in CHARMM (version 29). We are grateful to E. Guarnera and Dr. E. Paci for helpful discussions. We thank A. Widmer (Novartis Pharma, Basel) for providing the molecular modeling program WIT!P which was used for visual analysis of the trajectories. The simulations were performed on the Matterhorn Beowulf cluster at the Computing Center of the University of Zurich. We thank C. Bollinger and Dr. A. Godknecht for setting up the cluster and the Canton of Zurich for generous hardware support. This work was supported by the Swiss National Competence Center in Structural Biology (NCCR) and the Swiss National Science Foundation (Grant No. 31-64968.01 to A.C.).

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic Press, San Diego, 2002).
[2] B. J. Berne and J. E. Straub, Curr. Opin. Struct. Biol. **7**, 181 (1997).
[3] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[4] N. Rathore, T. A. Knotts IV, and J. J. de Pablo, J. Chem. Phys. **118**, 4285 (2003).
[5] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[6] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[7] C. D. Michele and F. Sciortino, Phys. Rev. E **65**, 051202 (2002).

[8] K. Sanbonmatsu and A. Garcia, Proteins: Struct., Funct., Genet. **46**, 225 (2002).
[9] A. E. Garcia and K. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).
[10] A. E. Garcia and K. Sanbonmatsu, Proteins: Struct., Funct., Genet. **42**, 345 (2001).
[11] R. Zhou, B. Berne, and R. Germain, Proc. Natl. Acad. Sci. U.S.A. **98**, 14931 (2001).
[12] A. E. Garcia and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **100**, 13898 (2003).
[13] J. W. Pitera and W. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[14] F. Rao and A. Caflisch, J. Chem. Phys. **119**, 4035 (2003).
[15] W. Im and C. L. Brooks, III, J. Mol. Biol. **337**, 513 (2004).
[16] C. M. Dobson, Trends Biochem. Sci. **24**, 329 (1999).
[17] M. F. Perutz, Trends Biochem. Sci. **24**, 58 (1999).
[18] C. Blake and L. Serpell, Structure (London) **4**, 989 (1996).
[19] S. B. Malinchik, H. Inouye, K. E. Szumowski, and D. A. Kirschner, Biophys. J. **74**, 537 (1998).
[20] M. Bucciantini, E. Giannoni, F. Chiti *et al.*, Nature (London) **416**, 507 (2002).
[21] M. Balbirnie, R. Grothe, and D. Eisenberg, Proc. Natl. Acad. Sci. U.S.A. **98**, 2375 (2001).
[22] R. A. Broglia, G. Tiana, S. Pasquali, H. E. Roman, and E. Vigezzi, Proc. Natl. Acad. Sci. U.S.A. **95**, 12930 (1998).
[23] D. Bratko and H. W. Blanch, J. Chem. Phys. **118**, 5185 (2003).
[24] G. Giugliarelli, C. Micheletti, J. R. Banavar, and A. Maritan, J. Chem. Phys. **113**, 5072 (2000).
[25] P. M. Harrison, H. S. Chan, S. B. Prusiner, and F. E. Cohen, J. Mol. Biol. **286**, 593 (1999).
[26] A. V. Smith and C. K. Hall, J. Mol. Biol. **312**, 187 (2001).
[27] B. Vekhter and R. S. Berry, J. Chem. Phys. **110**, 2195 (1999).
[28] M. Friedel and J. E. Shea, J. Chem. Phys. **120**, 5809 (2004).
[29] A. Fernandez and M. Boland, FEBS Lett. **529**, 298 (2002).
[30] J. Gsponer, U. Habertür, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **100**, 5154 (2003).
[31] E. Paci, J. Gsponer, X. Salvatella, and M. Vendruscolo, J. Mol. Biol. **340**, 555 (2004).
[32] F. Massi, J. W. Peng, J. P. Lee, and J. E. Straub, Biophys. J. **80**, 31 (2001).
[33] B. Ma and R. Nussinov, Protein Sci. **11**, 2335 (2002).
[34] B. Ma and R. Nussinov, Proc. Natl. Acad. Sci. U.S.A. **99**, 14126 (2002).
[35] D. Klimov and D. Thirumalai, Structure (London) **11**, 295 (2003).
[36] G. Tiana, F. Simona, R. A. Broglia, and G. Colombo, J. Chem. Phys. **120**, 8307 (2004).
[37] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Genet. **46**, 24 (2002).
[38] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[39] E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
[40] W. Hasel, T. F. Hendrickson, and W. C. Still, "A Rapid Approximation to the Solvent Accessible Surface Arteas of Atoms," Tetrahedron Computer Methodology (Pergamon, New York, 1998), Vol. **1**, No. 2, pp. 103–116.
[41] T. Lazaridis and M. Karplus, Proteins: Struct., Funct., Genet. **35**, 133 (1999).
[42] A. Hiltpold, P. Ferrara, J. Gsponer, and A. Caflisch, J. Phys. Chem. B **104**, 10080 (2000).
[43] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[44] P. Ferrara and A. Caflisch, J. Mol. Biol. **306**, 837 (2001).
[45] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **99**, 6719 (2002).
[46] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).
[47] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic Press, Oxford, 1990).
[48] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[49] M. E. Rose, *Elementary Theory of Angular Momentum* (Wiley, New York, 1957).
[50] C. Zannoni, *The Molecular Physics of Liquid Crystals* (Academic, London, 1979), Chap. 3.
[51] S. Chandrasekhar, *Liquid Crystals* (Cambridge University Press, Cambridge, England, 1992).
[52] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed. (Oxford University Press, Oxford, 1993).
[53] C. Zannoni, J. Mater. Chem. **11**, 2637 (2001).
[54] R. Berardi, L. Muccioli, and C. Zannoni, ChemPhysChem **5**, 104 (2004).
[55] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford Science, Oxford, UK, 1987).
[56] M. F. Perutz, T. Johnson, M. Suzuky, and J. T. Finch, Proc. Natl. Acad. Sci. U.S.A. **91**, 5355 (1994).
[57] M. F. Perutz, B. J. Pope, D. Owen, E. E. Wanker, and E. Scherzinger, Proc. Natl. Acad. Sci. U.S.A. **99**, 5596 (2002).
[58] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch (unpublished).
[59] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[60] M. K. Fenwick and F. A. Escobedo, J. Chem. Phys. **119**, 11998 (2003).
[61] A. Mitsutake, Y. Sugita, and Y. Okamoto, J. Chem. Phys. **118**, 6676 (2003).
[62] T. R. Serio, A. G. Cashikar, A. S. Kowal, G. J. Sawicki, J. J. Moslehi, L. Serpell, M. F. Arnsdorf, and S. L. Lindquist, Science **289**, 1317 (2000).
[63] T. P. Doerr, D. Herman, H. Mathur, and P. L. Taylor, Europhys. Lett. **59**, 398 (2002).
[64] Cecchini *et al.* (unpublished).
[65] M. Lopez de la Paz and L. Serrano, Proc. Natl. Acad. Sci. U.S.A. **101**, 87 (2004).
[66] W. DeLano, *The* PYMOL *Molecular Graphics System* (DeLano Scientific, San Carlos, CA, 2002).

Conclusions

Three new methods and their application for the study of the energy landscapes for folding have been presented.

A network based approach was introduced as a new framework for the study of the free-energy landscape sampled by molecular dynamics (MD) simulations of reversible folding. It was shown that, in contrast to the usual two-state scheme for folding [1], the denatured state of the structured peptide beta3s is highly heterogeneous with the presence of high entropy/high enthalpy basins as well as low entropy/low enthalpy basins [2]. Within the network framework, the states of the peptide are defined as the "communities" of the network [3].

A very fast method to characterize the transition state ensemble (TSE) was presented [4]. The folding probability p_{fold} is computed along the MD trajectory without requiring any additional simulation. The method is applied for the estimation of the TSE of a large set of mutants of beta3s and the computation of the $\Phi-$values [5].

Finally, replica exchange simulations (REM) were run and compared to classical constant temperature MD simulations [6, 7]. REM simulations have shown to correctly sample the energy landscape at physiological relevant temperatures which is impossible with constant temperature MD simulations.

Many questions remain open. A step forward for the research in the field would be to find a complete and consistent treatment of the errors in the study of the energy landscapes for folding. The overall picture presented in this thesis seems to be robust but quantitative criteria for the convergence and the confidence of the results are needed. Errors emerge from many aspects in the generation and analysis of an energy landscape and can be roughly summarized as: (1) Accuracy of the force field and solvation model, (2) incomplete sampling of the energy landscape, (3) definition of the states.

The last two points are tightly connected and the latter is partially a consequence of the former. Only a reliable and complete sampling of the landscape, hence of the different phases of the system, can lead to an accurate definition of the states. New simulation protocols, like REM, have shown to be of great value for this problem. However, they cannot be used to study folding pathways and the TSE because they don't give realistic kinetics. At this stage,

researchers are already trying to reconcile the REM protocol with the real dynamics of the system [8] using the network concepts introduced in this thesis. Such improvements in the use of REM and the network framework will be of great help to enhance our understandings in the mechanisms that govern protein folding.

Bibliography

[1] K.A. Dill, S. Bromberg, K.Z. Yue, K.M. Fiebig, D.P. Yee, P.D. Thomas, and H.S. Chan. Principles of protein-folding - a perspective from simple exact models. *Protein Science*, 4:561–602, Apr 1995.

[2] F. Rao and A. Caflisch. The protein folding network. *Journal of Molecular Biology*, 342:299–306, 2004.

[3] S. Muff, F. Rao, and A. Caflisch. Validation of network clusterizations. *cond-mat/0503252*, 2005.

[4] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch. Estimation of protein folding probability from equilibrium simulations. *Journal of Chemical Physics*, 122:184901, 2005.

[5] G. Settanni, F. Rao, and A. Caflisch. Φ-Value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. USA*, 102:628–633, 2005.

[6] F. Rao and A. Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *Journal of Chemical Physics*, 119:4035–4042, 2003.

[7] M Cecchini, F Rao, M Seeber, and A Caflisch. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *Journal of Chemical Physics*, 121:10748–10756, Dec 2004.

[8] Andrec M., Felts A.K., Gallicchio E., and Levy R.M. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. USA*, 102:6801, 2005.

List of Figures

1.1 **Funneled energy landscape for folding.** The overall shape of the surface is funnel-like with an energetic bias towards the native state. However, many local minima can arise as a consequence of competing chain interactions and show up as basins of attraction. Kinetics and thermodynamics of folding are strongly influenced by the presence of such heterogeneous non-native basins [12]. 2

1.2 **Entropic stabilization.** Despite an unfavorable internal energy, partially structured conformations can be stabilized entropically as in the case of some helical conformations of beta3s. 4

1.3 **Free-energy projections on order parameters.** In the case of beta3s, the fraction of native contacts doesn't necessarily identify structurally and kinetically homogeneous conformations. In the first, second and third column, conformations with $\sim 30\%$, $\sim 60\%$ and $\sim 90\%$ (native state) of native contacts Q are shown, respectively. The projected free-energy shows no evidence of the structurally and kinetically heterogeneity found in the denatured state of beta3s (see text). 5

1.4 **Beta3s free-energy landscape network.** Nodes and links are the conformations and the transitions between them, respectively. Red nodes represents the native basin. Node size is inversely proportional to the free-energy of the conformation. . . 8